TU Delft

Delft University of Technology

Document Version
Final published version

Citation (APA)
Ludwig, S., Aschenbrenner, D., Scharle, M., Klessig, H., Karrenbauer, M., Wu, H., Taghouti, M., Lozano, P., Schotten, H. D., & Fitzek, F. H. P. (2022). Reference Network and Localization Architecture for Smart Manufacturing Based on 5G. In M. Valle, D. Lehmhus, C. Gianoglio, E. Ragusa, L. Seminara, S. Bosse, A. Ibrahim, & K.-D. Thoben (Eds.), *Advances in System-Integrated Intelligence - Proceedings of the 6th International Conference on System-Integrated Intelligence SysInt 2022, Genova, Italy* (pp. 470-479). (Lecture Notes in Networks and Systems; Vol. 546 LNNS). Springer. https://doi.org/10.1007/978-3-031-16281-7_44

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

Lecture Notes in Networks and Systems 546

Maurizio Valle · Dirk Lehmhus · Christian Gianoglio · Edoardo Ragusa · Lucia Seminara · Stefan Bosse · Ali Ibrahim · Klaus-Dieter Thoben *Editors*

Advances in System-Integrated Intelligence

Proceedings of the 6th International Conference on System-Integrated Intelligence (SysInt 2022), September 7–9, 2022, Genova, Italy

Springer

Lecture Notes in Networks and Systems

Volume 546

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Advisory Editors

Fernando Gomide, Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil

Okyay Kaynak, Department of Electrical and Electronic Engineering, Bogazici University, Istanbul, Turkey

Derong Liu, Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA
Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Marios M. Polycarpou, Department of Electrical and Computer Engineering, KIOS Research Center for Intelligent Systems and Networks, University of Cyprus, Nicosia, Cyprus

Imre J. Rudas, Óbuda University, Budapest, Hungary

Jun Wang, Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

More information about this series at https://link.springer.com/bookseries/15179

Maurizio Valle · Dirk Lehmhus ·
Christian Gianoglio · Edoardo Ragusa ·
Lucia Seminara · Stefan Bosse ·
Ali Ibrahim · Klaus-Dieter Thoben
Editors

Advances in System-Integrated Intelligence

Proceedings of the 6th International Conference on System-Integrated Intelligence (SysInt 2022), September 7–9, 2022, Genova, Italy

Springer

Editors
Maurizio Valle
Department of Electrical, Electronic and Telecommunication Engineering and Naval Architecture (DITEN)
University of Genova
Genoa, Italy

Christian Gianoglio
Department of Electrical, Electronic and Telecommunication Engineering and Naval Architecture (DITEN)
University of Genova
Genoa, Italy

Lucia Seminara
Department of Electrical, Electronic and Telecommunication Engineering and Naval Architecture (DITEN)
University of Genova
Genoa, Italy

Ali Ibrahim
Department of Electrical and Electronics Engineering (EEE)
Lebanese International University (LIU)
Beirut, Lebanon

Dirk Lehmhus
Fraunhofer Institute for Manufacturing Technology and Advanced Materials
Bremen, Germany

Edoardo Ragusa
Department of Electrical, Electronic and Telecommunication Engineering and Naval Architecture (DITEN)
University of Genova
Genoa, Italy

Stefan Bosse
Computer Science and Mathematics
University of Bremen
Bremen, Germany

Klaus-Dieter Thoben
BIBA
University of Bremen
Bremen, Germany

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-031-16280-0 ISBN 978-3-031-16281-7 (eBook)
https://doi.org/10.1007/978-3-031-16281-7

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The International Conference on System-Integrated Intelligence, SysInt, celebrated its tenth anniversary with its 6th edition in 2022. Although SysInt has been conceived as an international event from the outset, this edition marked the first time the conference was held outside Germany. It moved to the bustling city of Genova with its formidable past and outstanding present, exemplified by its excellent reputation in scientific areas closely related to the main SysInt topics.

The hallmark of the SysInt conference series has always been to connect several research disciplines related to system-integrated intelligence, to deliberately provide food for thoughts crossing the individual boundaries of these fields, rather than promoting mutual exclusion. With its peculiar interdisciplinary approach, since its beginnings in 2012, SysInt has brought together scientists working on the fundamentals of smart products systems, with others dedicating to the application of such solutions. In terms of the latter, as in previous years, SysInt maintained its strong link to research in smart factories and logistic systems, often tagged with the term Industry 4.0. Beyond this, further applications covered by the event included soft robotics, human–machine interaction (HMI), pervasive and ubiquitous computing and not the least structural health monitoring (SHM). Enabling technologies included artificial intelligence (AI), cognitive systems as well as sensors and sensor systems. Most of these topics have been covered over several editions of the event and thus ensure that the SysInt conference proceedings can be considered as a valuable reference to discover development trends in such fields of research. This approach has proven its value over the decade of the conference existence. It has been supported by a continued interest in smart systems that has never been lost since 2012.

In the SysInt perspective, smart systems can be any stand-alone or product-integrated device which incorporate functionalities like sensing, actuation and data analysis, thus mediating between virtual and real environments, as well as superordinate systems made up of several such elements. Enabling technologies for smart systems can therefore be both hard- and software-related. Among the former are approaches required to synthesize the systems themselves or their components, like microsystem technologies, as well as methods to integrate them with arbitrary

products. Typical subsystems include sensors, actuators, data evaluation capabilities, energy supply and communication facilities. On the software side, communication protocols and energy management approaches as well as data evaluation strategies are of primary importance. A culmination point of combined efforts of this kind is material-integrated intelligent systems, which assume that functionality moves away from the product level toward becoming an inherent characteristic of the material. Challenges associated with smart systems in general are multiplied here, as resource efficiency becomes a primary concern.

Application scenarios for stand-alone as well as networked smart systems range from structural health monitoring to robotics, facilitation of human–machine interaction, smart manufacturing, the Internet of Things (IoT) or broader concepts like smart cities or pervasive and ubiquitous computing.

Thematically, the conference stays true to its origins, with many topics recurring from previous events both in the applied and in the fundamental fields of research. However, the general structuring has been adapted to include more special sessions than ever to accompany the main focus areas. This approach allows highlighting areas of specific relevance or timeliness, as well as providing a stage for bringing together renowned scientists as invited speakers and young researchers.

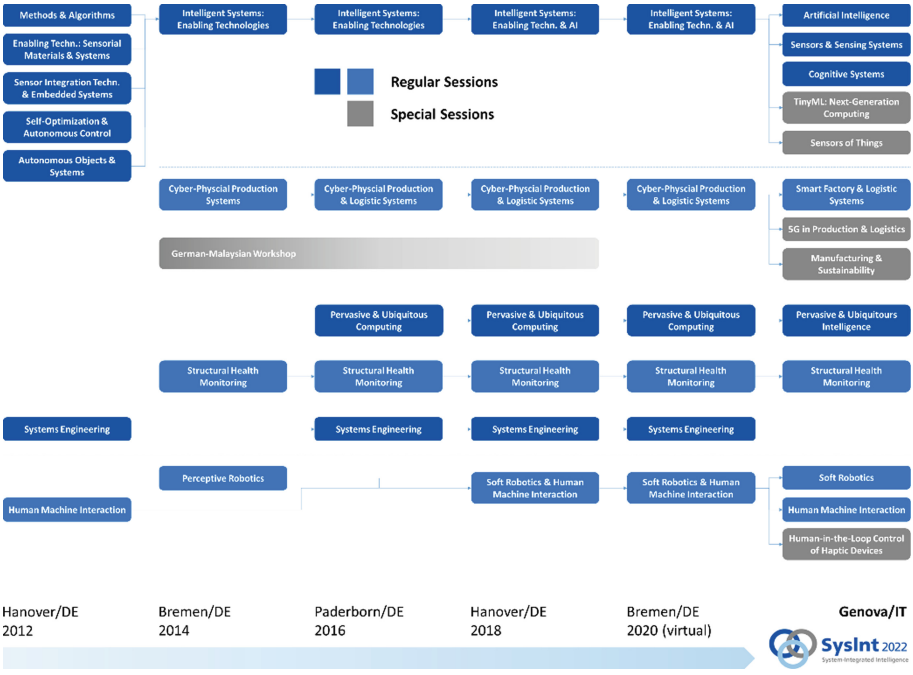


Fig. 1. Thematic areas of previous and current SysInt conferences, showing continuity of topics as well as new specializations.

Figure 1 illustrates the thematic width of the SysInt conference series, its interdisciplinary nature and the continuity of specific scientific strands. These date back to the very first SysInt conference held in Hanover, Germany, in 2012 based on an initiative of Berend Denkena, University of Hanover, Bernd Scholz-Reiter, University of Bremen, and Jürgen Gausemeyer, University of Paderborn. Furthermore, it highlights the changes that occurred over time. Most prominent is the aforementioned splitting of some of the long-standing topics into focused special sessions. Especially with respect to enabling technologies, the prevalence of AI-related topics is noteworthy—it reflects the increasing interest in such technologies, which has started to penetrate almost all areas of research in recent years. AI-related topics address a wide range of methods and algorithms including data-driven modeling and agent-based data processing and planning. The main goal is to achieve increased robustness and providing resilience by self-capabilities, e.g., self-organization and self-adaptivity. The new methods do not only address machines but do include humans as well, showing a current shift from Internet of Things to Internet of Humans architectures.

SysInt 2022 received a total of 72 full paper submissions. We considered this a very reasonable result given that (a) we were just recovering from the COVID-19 pandemic and only starting to re-establish physical events of this kind, (b) by the registration deadline it was far from clear what the situation would be in this respect toward the date of the conference in early September and this certainly discouraged a number of registrations and (c) the submission procedure was significantly altered, with authors being required to submit full papers rather than short abstracts for the first time in the course of the series.

Responsibility for the peer review process was handed to the track and special session chairs, who invited reviewers based on their personal contacts, while also drawing on the pool of organizing and scientific committee members. As a result of this distributed approach, peer review proved to be smooth and timely. The minimum number of requested reviews for each paper was three. Final decisions were taken based on a meta-review performed by the organizing committee. Out of the 72 submissions, 53 were of sufficient quality for direct acceptance, requiring only minor revision. The remaining ones were either sent back to the authors for major revision or rejected (2). This translates to an overall acceptance ratio of 97.2 percent, which is justified by the high quality of the contributions received.

The conference program has been further enriched by some internationally renowned keynote and invited speakers inspiring interesting discussion during the main scientific tracks as well as special sessions.

6th International Conference on System-Integrated Intelligence (SysInt2022)

General Chair

Maurizio Valle	University of Genoa, Italy

Organizing Committee

Stefan Bosse	University of Bremen, Germany
Christian Gianoglio	University of Genoa, Italy
Aleksandra Himstedt	Bremer Institut BIBA, Germany
Ali Ibrahim	Lebanese International University, Lebanon
Dirk Lehmhus	Fraunhofer IFAM, Germany
Edoardo Ragusa	University of Genoa, Italy
Lucia Seminara	University of Genoa, Italy
Silvia Ferrari	Pragma Congressi srl, Italy

International Scientific Committee

Giovanni Berselli	University of Genoa, Italy
Stefan Bosse	University of Bremen, Germany
Agostino Bruzzone	University of Genoa, Italy
Alessandro Bruzzone	University of Genoa, Italy
Lorenzo Capineri	University of Florence, Italy
Fu-Kuo Chang	Stanford University, USA
Hussein Chible	Lebanese University, Lebanon
Stephen Chih-Yang Lu	University of Southern California, USA
Nikolaus Correll	University Colorado Boulder, USA
Doriana D'Addone	University of Naples Federico II, Italy
Berend Denkena	University of Hanover, Germany
Christian Gianoglio	University of Genoa, Italy
Nikhil Gupta	New York University, USA
Fotis Kopsaftopoulos	Rensselaer Polytechnic Institute, USA

Ali Ibrahim	Lebanese International University, Lebanon
Dirk Lehmhus	Fraunhofer IFAM, Germany
Björn Lüssem	University of Bremen, Germany
Stefano Mariani	Polytechnic University of Milan, Italy
Yasser Mohanna	Lebanese University, Lebanon
Hassan Mostafa	Cairo University, Egypt
Changjae Oh	Queen Mary University London, UK
Max Pfingsthorn	University of Oldenburg, Germany
Pier Paolo Pompa	Italian Institute of Technology, Italy
Edoardo Ragusa	University of Genoa, Italy
Lucia Seminara	University of Genoa, Italy
Michael Sinapius	TU Braunschweig, Germany
Roberto Teti	University of Naples Federico II, Italy
Klaus-Dieter Thoben	University of Bremen, Germany
Flavio Tonelli	University of Genoa, Italy
Lin Wang	Queen Mary University London, UK
Thorsten Wuest	West Virginia University, USA

Technical Program Chairs

Christian Gianoglio	University of Genoa, Italy
Edoardo Ragusa	University of Genoa, Italy

Track and Special Session Chairs

Stefan Bosse	University of Bremen, Germany
Christian Gianoglio	University of Genoa, Italy
Edoardo Ragusa	University of Genoa, Italy
Björn Lüssem	IMSAS, University of Bremen, Germany
Flavio Tonelli	University of Genoa, Italy
Klaus-Dieter Thoben	BIBA University of Bremen, Germany
Lorenzo Capineri	University of Florence, Italy
Fotis Kopsaftopoulos	Rensselaer Polytechnic, Troy, NY, USA
Giovanni Berselli	University of Genoa, Italy
Changjae Oh	Queen Mary University, UK
Lin Wan	Queen Mary University, UK
Lucia Seminara	University of Genoa, Italy
Max Pfingsthorn	University of Oldenburg, Germany
Ali Ibrahim	Lebanese International University, Lebanon

Organizers

Università di Genova

Bremer Institut für Produktion und Logistik (BIBA)

Fraunhofer-Institut für Fertigungstechnik und Angewandte Materialforschung (IFAM)

Cosmic Lab

Dipartimento di Ingegneria Navale, Elettrica, Elettronica e delle Telecomunicazioni (DITEN)

LogDynamics Lab

Sponsors

CIRP

IEEE Sensor Council

The American Society of Mechanical Engineers (ASME)

Contents

Artificial Intelligence

Towards Challenges and Proposals for Integrating and Using Machine Learning Methods in Production Environments 3
Thimo Florian Schindler, Dennis Bode, and Klaus-Dieter Thoben

Autonomous Driving Based on Imitation and Active Inference 13
Sheida Nozari, Ali Krayani, Pablo Marin, Lucio Marcenaro, David Martin, and Carlo Regazzoni

Machine Learning Based Reconstruction of Process Forces 23
Berend Denkena, Heinrich Klemme, and Dennis Stoppel

A Novel Rule-Based Modeling and Control Approach for the Optimization of Complex Water Distribution Networks 33
Enrico Ferrari, Damiano Verda, Nicolò Pinna, and Marco Muselli

Graph-Based Segmentation and Markov Random Field for Covid-19 Infection in Lung CT Volumes 43
Giulia Iaconi, Federica Ferraro, Marco Balletto, David Solarna, Marco Trombini, Gabriele Moser, and Silvana Dellepiane

Image Based Classification of Methods-Time Measurement Operations in Assembly Using Recurrent Neuronal Networks 53
Patrick Rueckert, Katrin Birgy, and Kirsten Tracht

Pervasive and Ubiquitous Intelligence

FPGA-Based Road Crack Detection Using Deep Learning 65
Lorenzo Canese, Gian Carlo Cardarilli, Luca Di Nunzio, Rocco Fazzolari, Marco Re, and Sergio Spanò

Simple Non Regressive Informed Machine Learning Model for Prescriptive Maintenance of Track Circuits in a Subway Environment 74
Andrea Garrone, Simone Minisi, Luca Oneto, Carlo Dambra, Marco Borinato, Paolo Sanetti, Giulia Vignola, Federico Papa, Nadia Mazzino, and Davide Anguita

Embedded Implementation of an Algorithm for Online Inertia Estimation in Power Grids 84
Alessandro Ravera, Valentina Baruzzi, Matteo Lodi, Alberto Oliveri, and Marco Storace

Random Weights Neural Network for Low-Cost Readout of Colorimetric Reactions: Accurate Detection of Antioxidant Levels 95
Edoardo Ragusa, Valentina Mastronardi, Deborah Pedone, Mauro Moglianetti, Pier Paolo Pompa, Rodolfo Zunino, and Paolo Gastaldo

Resource-Constrained Implementation of Deep Learning Algorithms for Dynamic Touch Modality Classification 105
Haydar Al Haj Ali, Christian Gianoglio, Ali Ibrahim, and Maurizio Valle

Human Recognition for Resource-Constrained Mobile Robot Applied to Covid-19 Disinfection 116
Andrea Mattia Garavagno, Daniele Leonardis, and Antonio Frisoli

Data-Driven Methods for Aviation Safety: From Data to Knowledge 126
Irene Buselli, Luca Oneto, Carlo Dambra, Christian Verdonk Gallego, and Miguel Garcia Martinez

Design and Deployment of an Efficient Landing Pad Detector 137
Andrea Albanese, Tommaso Taccioli, Tommaso Apicella, Davide Brunelli, and Edoardo Ragusa

Towards a Trade-off Between Accuracy and Computational Cost for Embedded Systems: A Tactile Sensing System for Object Classification 148
Youssef Amin, Christian Gianoglio, and Maurizio Valle

An Optimized Heart Rate Detection System Based on Low-Power Microcontroller Platforms for Biosignal Processing 160
Benedetta Mazzoni, Giuseppe Tagliavini, Luca Benini, and Simone Benatti

Sensors and Sensing Systems

A Non-Hilbertian Inversion Technique for the Diagnosis of Faulty Elements in Antenna Arrays 173
Valentina Schenone, Alessandro Fedeli, Claudio Estatico, Matteo Pastorino, and Andrea Randazzo

A Passive, Wireless Sensor Node for Material-Integrated Strain and Temperature Measurements in Glass Fiber Reinforced Composites 182
Lukas Bertram, Michael Brink, Klaus-Dieter Thoben, and Walter Lang

Multi-camera Metrology System for Shape and Position Correction of Large Fuselage Components in Aircraft Assembly 194
Daniel Valencia, Jörg Wollnack, Sanjay Kamath, and Leander Brieskorn

Management of Research Field Data Within the Concept of Digital Twin 205
Hauke Dierend, Osman Altun, Iryna Mozgova, and Roland Lachmayer

Feed-Forward SNN for Touch Modality Prediction 215
Ali Dabbous, Ali Ibrahim, and Maurizio Valle

Smart Factory and Logistic Systems

Ansaldo Energia Progetto LHP (OR6.3) 225
Roberto Mosca, Marco Mosca, Roberto Revetria, Saverio Pagano, and Federico Briatore

An Application of Engineering 4.0 to Hospitalized Patients 235
Roberto Mosca, Marco Mosca, Roberto Revetria, Fabio Currò, and Federico Briatore

A DT-Based System for Predicting Process Behavior 245
Bowen Qi and HongSeok Park

Optimal Robot Workpiece Placement for Maximized Repeatability 252
Jan Baumgärtner, Philipp Gönnheimer, and Jürgen Fleischer

Enhancing Vendor Managed Inventory with the Application of Blockchain Technology 262
Santhosh Ganesan, Hendro Wicaksono, and Omid Fatahi Valilai

Design and Development of a Tomato Picking Soft Robotic Gripper with a Separator and Mechanical Iris Based Pedicel Cutting Mechanism 276
Shahid Ansari and Bishakh Bhattacharya

The Impact of Intelligent Objects on Quality 4.0 287
Sergio Salimbeni and Andrés Redchuk

Mutual Resiliency and Lean Analysis for Supply Chain Management in a COVID-19 Ventilator Production Case Study 299
Alan Caraveo, Saikiran Krishnamoorthy, Jonah Marie Lawas, Parthasarathy Periaswamy, Majid Sodachi, and Omid Fatahi Valilai

Fighting Hospital Infections with Engineering 4.0 310
Roberto Mosca, Marco Mosca, Roberto Revetria, Fabio Currò, and Federico Briatore

Cyber-Physical Equipment as a Service 318
Gustavo Sanchez, Giancarlo Bo, Fabrizio Cardinali, and Flavio Tonelli

Structural Health Monitoring

A Sensing System for Defects Detection in Reinforced Aluminium Panels by Ultrasonic Guided Waves Sensors 331
Lorenzo Capineri, Andrea Bulletti, Cosimo Damiani, and Luca Bergamaschi

A Tiny Machine Learning Approach to the Edge Localization of Acoustic Sources via Convolutional Neural Networks 340
Federica Zonzini, Giacomo Donati, and Luca De Marchi

Integrating Electronic Components, Sensors and Actuators in Cast Metal Components: An Overview of the State of the Art 350
Dirk Lehmhus, Thomas Rahn, Christoph Pille, and Matthias Busse

Fast Temperature-Compensated Method for Damage Detection and Structural Health Monitoring with Guided Ultrasonic Waves and Embedded Systems 362
Christoph Polle, Stefan Bosse, Michael Koerdt, Björn Maack, and Axel S. Herrmann

A Study on a Novel Impact Test for SHM of CFRP Based on Screwdriver Bit Launch and Electret Microphone 379
Cristiano Soares Junior, Paulo R. Aguiar, Pedro O. C. Junior, Reinaldo Gotz, Paulo M. C. Monson, and Alessandro Roger Rodrigues

Optimization of Non-destructive Damage Detection of Hidden Damages in Fiber Metal Laminates Using X-ray Tomography and Machine Learning Algorithms 387
Chirag Shah, Stefan Bosse, Carolin Zinn, and Axel von Hehl

Human Machine Interaction

Time Reduction in Online Programming – An Approach to Hand Guided Teaching for Small Batch Robot Machining 405
Marten Stepputat, Florian Beuss, Jan Sender, and Wilko Fluegge

Ontology-Based Documentation of Quality Assurance Measures Using the Example of a Visual Inspection 415
Tatyana Sheveleva, Kevin Herrmann, Max Leo Wawer, Christoph Kahra, Florian Nürnberger, Oliver Koepler, Iryna Mozgova, Roland Lachmayer, and Sören Auer

Controlling Decisions by Head Electrical Signals 425
Enrico Zero, Alessandro Bozzi, Simone Graffione, and Roberto Sacile

Tactile-Based Human-Robot Collaboration: A Performance Analysis 437
Francesco Grella, Roberto Canale, Francesco Giovinazzo, Alessandro Albini, and Giorgio Cannata

Applications and Challenges in the Use of 5G in Production and Logistics

Wireless Communication Technologies in Smart Factories 449
Dieter Lutzmayr and Manfred Pauritsch

Concept of a 5G Hybrid Wireless Campus Network as Testbed for Industrial Applications 458
Christoph Cammin, Thomas Doebbert, Bettina Solzbacher, and Gerd Scholl

Reference Network and Localization Architecture for Smart Manufacturing Based on 5G 470
Stephan Ludwig, Doris Aschenbrenner, Marvin Scharle, Henrik Klessig, Michael Karrenbauer, Huanzhuo Wu, Maroua Taghouti, Pedro Lozano, Hans D. Schotten, and Frank H. P. Fitzek

Survey on Usage of 5G Campus Networks in Intralogistics 480
Lara Nehrke, Simone Neumann, Alexandra Cieslak, and Alice Kirchheim

Concept for a Revolution of Public Transport 489
Wolfgang Echelmeyer, Tuan Nguyen, and Mert Mete

Human-in-the-Loop Control of Haptic Devices: Now and the Future

An Intelligent System for Human Intent and Environment Detection Through Tactile Data 497
Gianluca Laudante and Salvatore Pirozzi

Design of a Wearable Haptic Device to Mediate Affective Touch with a Matrix of Linear Actuators 507
Nikolas Ferguson, Mehmet Ege Cansev, Anany Dwivedi, and Philipp Beckerle

Motor-Unit Ordering of Blindly-Separated Surface-EMG Signals for Gesture Recognition 518
Mattia Orlandi, Marcello Zanghieri, Davide Schiavone, Elisa Donati, Francesco Conti, and Simone Benatti

The Effect of Cognitive Load on Electrotactile Communication via a Multi-pad Electrode 530
Fabricio Ariel Jure, Erika Geraldina Spaich, Laura Petrini, and Strahinja Dosen

TinyML: The Quest for Next-Generation Computing

Inter-Operability of Compression Techniques for Efficient Deployment of CNNs on Microcontrollers 543
Hamoud Younes, Hugo Le Blevec, Mathieu Léonardon, and Vincent Gripon

Energy Consumption Analysis of Pruned Semantic Segmentation Networks on an Embedded GPU 553
Hugo Tessier, Vincent Gripon, Mathieu Léonardon, Matthieu Arzel, David Bertrand, and Thomas Hannagan

A Tiny CNN for Embedded Electronic Skin Systems 564
Fouad Sakr, Hamoud Younes, Joseph Doyle, Francesco Bellotti, Alessandro De Gloria, and Riccardo Berta

Detecting Ice on Wind Turbine Rotor Blades: Towards Deep Transfer Learning for Image Data 574
Maria Teresa Alvela Nieto, Hannes Gelbhardt, Jan-Hendrik Ohlendorf, and Klaus-Dieter Thoben

Towards Real-Time Human Detection in Maritime Environment Using Embedded Deep Learning 583
Mostafa Rizk, Fatima Slim, Amer Baghdadi, and Jean-Philippe Diguet

Advancement in Manufacturing and Sustainability

Data Driven Decision Making When Transitioning Towards a Modular Setup 597
Morten Skogstad Nielsen, Thomas Ditlev Brunoe, Ann-Louise Andersen, and Kjeld Nielsen

A Novel Method for Component Positioning in Thermoformed Electronics 607
Behnam Madadnia, Frederick Bossuyt, and Jan Vanfleteren

Towards Circular Production Systems: Outlining the Concept, Challenges and Future Research Directions 616
Filip Skärin, Carin Rösiö, and Ann-Louise Andersen

Facilitating Manufacturing System Development: Mapping Changeability Capabilities in Two Industrial Cases 626
Stefan Kjeldgaard, Rasmus Andersen, Alessia Napoleone, Thomas Ditlev Brunoe, and Ann-Louise Andersen

Transfactory: Towards a New Technology-Human Manufacturing Co-evolution Framework 636
Fabrizio Bracco, Alessandro Arturo Bruzzone, and Emanuele Carpanzano

Sustainability Comparison of a New Reconfigurable Machine Tool and a Conventional Milling Machine 646
Alessandro Arturo Bruzzone, Alessandra Ferrari, and Alessia Napoleone

Comparison of Machine Learning Models for Predictive Maintenance Applications 657
Alessia Lazzaro, Doriana Marilena D'Addona, and Massimo Merenda

Exploring Manufacturing System Development and the Use of Platforms to Reduce Time-to-Market 667
Morten Skogstad Nielsen, Ann-Louise Andersen, Thomas Ditlev Brunoe, Khaled Medini, and Kjeld Nielsen

Sensor of Things

Automated Tonic-Clonic Seizure Detection Using Random Forests and Spectral Analysis on Electroencephalography Data 679
Craig Stewart, Wai Keung Fung, Nazila Fough, and Radhakrishna Prabhu

An Investigation into Routing Protocols for Real-Time Sensing of Subsurface Oil Wells 689
Craig Stewart, Nazila Fough, and Radhakrishna Prabhu

Surface-Functionalized Multichannel Nanosensors and Machine Learning Analysis for Improved Sensitivity and Selectivity in Gas Sensing Applications 700
Luis Antonio Panes-Ruiz, Shirong Huang, Leif Riemenschneider, Alexander Croy, Bergoi Ibarlucea, and Gianaurelio Cuniberti

Design and Simulation of a Novel Low-Voltage RF MEMS Switch for Reconfigurable Antennas 708
Alaa Elshazly, Mario Mounir, Muhammad K. Khalaf, Faycal Saffih, Yasmine Elogail, and Hassan Mostafa

2T1M Neuromorphic Synapse with Pt-Hf-Ti Memristor Model 714
Heba Allah Gamal, Hassan Mostafa, and Ayman Haggag

Embedded Implementation of Signal Pre-processing for Tactile Sensing System 727
Moustafa Saleh, Yahya Abbass, and Maurizio Valle

Author Index 737

Artificial Intelligence

Towards Challenges and Proposals for Integrating and Using Machine Learning Methods in Production Environments

Thimo Florian Schindler(✉), Dennis Bode, and Klaus-Dieter Thoben

BIBA - Bremer Institut für Produktion und Logistik GmbH,
Hochschulring 20, 28359 Bremen, Germany
sth@biba.uni-bremen.de
https://biba.uni-bremen.de

Abstract. The integration of complex and innovative technologies into manufacturing processes poses various challenges. This paper comprehensively outlines the challenges that can occur when integrating and using machine learning (ML)-based methods in a production environment to assist in steering the process. The identified problems are considered regarding different categories of problems in process modelling. In addition, potential innovative technologies and methods are shown that can help to mitigate the challenges. Various applications and projects dealing with increasing energy efficiency through ML served as use cases to identify the challenges presented and possible suggestions to mitigate those same challenges.

Keywords: Manufacturing · Integration challenges · Machine learning · Data-driven modelling

1 Introduction

In the context of modern factory processes and advancing digitalisation in companies, it can be shown that the use of new technologies, such as machine learning methods, can exploit previously unused energy-saving potential in production and process parameters to increase energy efficiency [2]. The leading integrated models used in this paper are based on machine learning methods (ML). In comparison to an offline ML-based use and evaluation of the production processes, the integration of ML-services into production is challenging and can cause several problems. This paper describes the overarching general challenges of integrating and using ML-based methods and shows potential technical and methods approach to mitigate these challenges.

Upon its challenges, the ML integration concept is built up concerning CRISP-DM, DUCAR, SCRUM and constantly refers to the integration carried out in a production environment based on results from application-oriented research projects [10].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 3–12, 2023.
https://doi.org/10.1007/978-3-031-16281-7_1

The majority of these tools select suitable machine learning methods and prototype them in the correspondingly identified areas of the production and process environment. However, this paper focuses on the challenges encountered when integrating and using these techniques into production environments from research. Therefore in this paper, the possible challenges that come with the ML-service integration are identified based on several use-cases and potential approaches to mitigate these problems are presented. After all, the successful integration and use of machine learning methods in modern production environments can increase energy efficiency in companies' internal processes [2].

1.1 Motivation

Integrating machine learning concepts in manufacturing processes is only slowly finding its way into small and medium-sized enterprises [10]. One reason for this hesitation or the currently still low use of machine learning methods is that these methods are unfamiliar and unknown for small and medium-sized companies [6]. The resulting progress and optimisation of process control concerning the effort required to develop such systems are still novel.

Another reason why artificial intelligence has not yet found widespread application in small and medium-sized enterprises is that there is still a lack of systematic approaches to linking data science with the areas of company-specific production processes. This paper contributes to the widespread application of ML in industrial processes by proposing approaches to address the challenges mentioned in this article.

Furthermore, companies are faced with the challenge of collecting data in the context of digitalisation and providing already acquired data resources in big data to develop data-driven models. All these aspects lead to the problem that patterns in the data for process optimisation cannot be identified. In order to be able to solve these challenges, it is essential to begin modelling at the process capture stage and consider it holistically up to the process control and model integration stage. Because findings from current industry-related applications of machine learning methods show that even small process-parameter changes can lead to far-reaching improvements and increases in energy efficiency [2] and so in sustainability. In addition to these challenges, another hurdle for SMEs is that it is difficult to assess from the outset to what extent the technical and economic risks of integrating ML-based modelling methods can be classified. These risks will not be discussed in detail in this paper. However, these risks are also associated with the integration challenges described here. Therefore, it can be concluded that further steps are needed to ensure the successful and holistic integration and facing the upcoming challenges of machine learning methods into the production environment.

2 General Overview

As part of the funded projects KIPro, ErProPlus, vKBP and Tide2Use, prototype machine learning-based modules were developed in a research environment.

As an assistance system, these should improve the energy efficiency of production processes by providing suggestions for action to control the production systems and service processes. The necessary steps to integrate and use these modules include the five steps illustrated in Fig. 1.

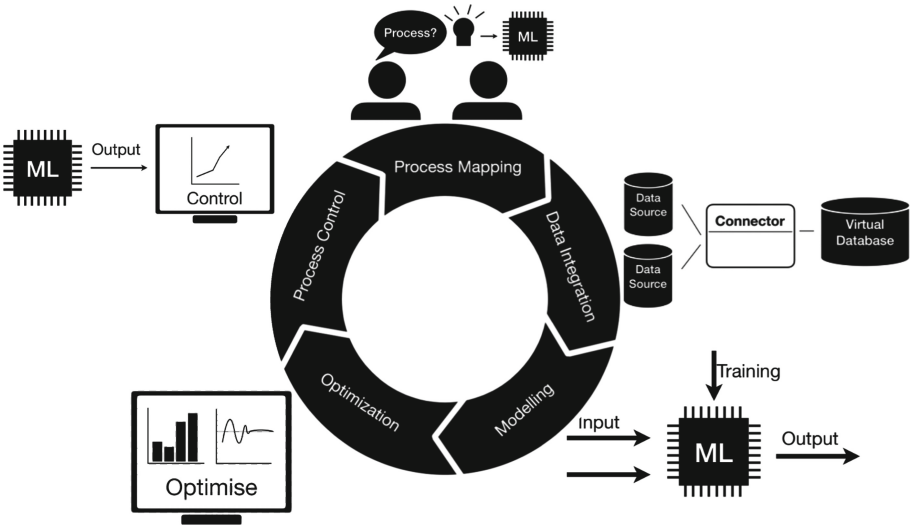


Fig. 1. Five steps of machine learning projects, based on CRSIP-DM [4]

When integrating and using machine learning-based methods into the production or service processes of the industrial partners, many difficulties arose as experience. References to similar issues from research support the experiences presented in this paper. The experiences with these are to be recorded and presented in this paper using a systematic approach. In addition, approaches to some of these problems are already presented and proposed, some of which have been evaluated in the application projects.

In Table 1 an overview of the occurring challenges when integrating and using ML-based methods in a production environment is presented. The order of the challenges identified in Table 1 is mapped to Fig. 1 for machine learning projects and the five steps shown there. The challenges will be discussed in detail in Sect. 3.1.

3 Identified Challenges and Proposals

3.1 Social Challenges and Human Factors

When implementing and integrating machine learning-based models in production systems, there are technical challenges and social and human aspects. The

Table 1. Identified challenges related to the research work on machine learning projects and their five steps

Nr.	Identified challenges	Description	ML-project phase
3.1	Social challenges and human factors	How social factors can influence machine learning integration	Process mapping
3.2	Sensors and data sources	Focus on data sources and sensor technology about associated challenges	Data integration
3.3	Computational and processing capacity	The difference in computing power between development and production system environments leads to specific challenges	Modelling
3.4	Software dependencies	Concomitant challenges due to networked and interdependent software applications	Modelling
3.5	Model availability and system failures	Continuous model availability is essential for successful machine learning integration	Optimisation
3.6	Adaptation of business processes and model adjustments	How process changes and model adjustments disrupt the accuracy of predictions	Process control

employees in the companies have a high degree of expert knowledge of the processes underlying the models. The employees usually accompany and implement the strategies in the companies for several years, if not decades.

It is often the case that the workforce has unfounded fears about their jobs due to insufficient knowledge about those methods. Due to the complexity and effort required for machine learning-based modelling and integration into the process environment.

In particular, neural networks also have the problem that it is difficult or even impossible to conclude their predictions. It is therefore elaborating to reconstruct decisions and calculations from these ML models. This makes it challenging to justify incorrect predictions to the employees who use the models in the production environment.

For the researchers or third parties not directly involved in the process, the challenge is to think their way into a complex, possibly multi-layered, unknown production or process environment to recognise sociological dependencies and system interrelationships. Often there is a unique and outwardly isolated understanding of the process among the employees, which may be unknown to a third party.

These are some of the challenges that make it unavoidable to increase the intrinsic motivation of the employees to embed a successful integration of ML-based models into the processes.

Suppose machine learning-based models are to find their way into modern production environments. In that case, employees must be involved right from the beginning of the process development phase of a project to reduce employees' fears early and make expert knowledge available.

Although artificial neural networks are trained to behave predictably, tracing back exactly how the model calculated its results is impossible. This leads to uncertainties about the model predictions and, in the event of incorrect predictions, can only be explained to the employees of a company with difficulty. To counteract this, direct and clear communication about the power and the disadvantages of such algorithms with the employees should be strived for.

Within the framework of proper monitoring, the operator must be shown transparently the results and the quality of the model calculations of the assistance systems to be integrated.

Furthermore, transparent and regular communication between employees and third parties creates a joint knowledge base about the underlying process. Training and questionnaires also help to improve common process understanding and communication [5]. It must also be made clear to employees what the goals of integrating ML models are.

3.2 Sensors and Data Sources

Machine-learned models require data and information about the processes they are designed to model. In a live or training scenario operation, it may later happen that specific sensors fail or produce falsy measures that are temporarily unavailable or are replaced by other sources [9]. When integrating machine learning into production plants, dealing with the models with this changing information base is another upcoming challenge.

One of these challenges is the ability to use redundant data sources so that an alternative source of information is available in a failure. Another important aspect is the possibility of flexibly retraining the models in the production environment. In this way, the models can also be reliably adapted to changing conditions in the database.

3.3 Computational and Processing Capacity

Machine learning methods, particularly the training of deep artificial neural networks, require high computing power to develop the models to be depicted with sufficient accuracy and in a reasonable amount of time [4]. The rapid increase in available and usable processing power, both on local devices and in decentralised cloud services, is responsible for the fact that machine learning methods have found their way into scientific computing and in production environments. For many companies, data protection reasons and internal company trade secrets play an essential role, which is why they decide against cloud computing as a first step. It is also interesting for SMEs that already have an internal computing infrastructure to use innovative ML algorithms in their corporate processes cost-effectively and with existing resources.

Without the technical computing power on which the system is based, the training of far more complex network structures, such as bi-directional recurrent neural networks, but also already correspondingly deep feedforward neural networks, is only possible to a limited extent.

By analogy with Moore's law, Agrawal et al. describe how a significant increase in computer processing power can affect the effort and cost of applying machine learning to many different technical domains. In summary, an improvement in data storage, data processing and data provision means a dramatic reduction in the cost of machine learning-based predictions [1].

Modern neural network structures can be trained with the help of efficient backpropagation algorithms. The undetermined weighting parameters can be adjusted for a sufficiently accurate approximation of the real-time process by feeding the prediction errors back into the networks. To solve scientific and technical problems, efficient and computationally powerful GPUs must be used [11]. When developing models with the help of ML methods, a large number of prototype models can already be created due to the significantly increased computing power on local terminals.

During the development and subsequent use of machine learning-based models, sufficient computing power is essential to process large amounts of data in an adequate amount of time and train and calculate the models on this data. Therefore, it is necessary to ensure that sufficient computing capacity is available both in the development and production system environments. Analysing the required performance within the model development helps estimate the servers' essential computing and storage resources in the productive system environment.

Tuli et al. show that it is advantageous for applications in near-real-time environments if the models are deployed on cloud computing platforms to implement live prediction and more accurate models. In addition, to rapidly enhance model development and forecast, the use of such novel energy-efficient edge systems can also reduce power consumption [12].

It should be noted that other applications also require computing resources. Artificial neural networks need high computing capacities, especially when training the networks, while general calculations after the training phase of the models are far less computationally intensive. This means that it must be ensured that the hardware can provide sufficient capacity, especially if the models need to be retrained.

Using containerised Docker environments, individual hardware components, such as RAM or graphics cards, can be used selectively and flexibly. Furthermore, computationally intensive tasks, such as the training of artificial neural networks, can be outsourced to cloud services. To flexibly use several physical computing clusters in parallel, Kubernetes can be used to orchestrate computationally intensive tasks and outsource them to the respective clusters and specifically allocate resources [8].

3.4 Software Dependencies

Particularly in research projects, initial ML-based forecasting models are often prototyped on local computing units. Suppose these generated models are integrated directly into an application or industrial environment. In that case, the challenge is that the industrial process's computing infrastructure has potential and sensitive differences on which the forecasting models have been developed, trained, and tested. The differences include the operating systems used and various software environments which are already installed that are incompatible with the software packages used in the prototype development. Specifically, some software applications have dependencies on definitive versions of other software libraries. If multiple programmes use the same software library with different versions, this can easily lead to their inability to work in parallel on one server. To counteract this, technologies must be used to encapsulate these prototype systems. This reduces the risk of failure and increases the robustness of the individual software components.

Furthermore, encapsulation of the systems leads to improved data and failure safety. Still, it challenges that communication between these services can be more complex and have inferior performance. The difficulty that encapsulation also causes is that the maintenance effort and access possibilities to log information are increased due to encapsulated applications.

Separated and isolated software systems, such as data acquisition and model calculation, must be easily migrated from the development environment to the application level.

A resulting failure of the systems caused by unavailable resources, can lead to an overall loss of the integrated machine learning-based models and their computational components. Insofar as development does not occur directly on the same server computing infrastructure, which is often the case in research projects, differences cannot be avoided.

Docker can be used for container virtualisation to isolate and encapsulate software applications. All the required packages of the programmes are contained within the individual container environments [7]. This means that different software programmes can use different package versions and work independently. The encapsulated containers can be easily transported after the actual model development and installed on others in the production system environment.

Furthermore, secure communication between the individual containers is guaranteed via ports and isolated network structures within the Docker environment. Standard programme outputs can be generated and monitored via log files on persistent connected data storage.

In addition, the individual containers can be allocated the corresponding computing resources and storage capacities of the system on which the container is based.

3.5 Model Availability and System Failures

Using ML methods directly integrated into production, model relationships are mapped based on many measured variables. If the ML model fulfils an essential

task such as controlling the production plant, a high level of availability of the model forecasts is critical. It has been shown that this availability is a challenging factor. On the one hand, data sources that the model requires can fail. On the other hand, the availability of the infrastructure and networks needed for the model query is not always given. The mentioned challenges were identified in several projects and are also mentioned in [3].

A possible approach to avoid these problems is dynamic service and data source monitoring, including a notification system when data is not available. An example application is the open source network monitoring system Zabbix, which can be used to monitor system services and network connections.

3.6 Adaptation of Business Processes and Model Adjustments

The challenge when integrating ML-based methods in production processes is the changes in the business process to be modelled. For example, new products or the exchange or modification of production facilities can create unique circumstances that were not included in the training of the machine-learned model. As a result, these models can no longer provide predictions for the current situation of the process.

To react to changes in the process and keep the model performance high, a continuous evaluation and monitoring of the metrics of the model is necessary. At the same time, there must be the possibility of retraining the models in the event of an increasing or excessive deviation. In this context, it is also essential to take transfer learning strategies into account to reduce the training effort [13].

4 Summary and Conclusion

The focus of this paper was to take a closer look at the challenges associated with integrating and using machine learning-based models into a production environment and to propose specific methods for solving those challenges. The paper showed that technical challenges could influence and hinder the integration of ML-based methods. Social factors can be effectively solved through education and communication between researchers and employees in the companies to implement improved integration and use of machine learning. Via modern software technologies, such as Docker containers or Kubernetes, complex software applications could be interconnected, developed, and integrated into different environments. The paper has focused on describing the challenges significantly discovered in the research projects and presenting proposals. Obviously, many additional challenges and hurdles still exist when digitising and integrating ML-based models permanently in use in a production environment. This also involves investigating the extent to which the effort required to develop, maintain and service ML-based methods compares to conventional methods for modelling and increasing energy efficiency in industrial processes. Also in areas where exact traceability of the modelling results is necessary, it may be disadvantageous to use black-box models like neural networks. In the context of future work, it

should be investigated how far the findings made in this paper would behave in integrating ML-based methods in cloud computing systems for production processes. Based on the knowledge gained in the research projects about possible approaches for integrating and using ML, the occurring challenges could be mitigated.

Acknowledgement. The authors would like to thank their project partners and the anonymous reviewers for their valuable input. The "Tide2Use" project (19H18004) is funded by the German Federal Ministry of Transport and Digital Infrastructure (BMDV) in the "Innovative Port Technologies" (IHATEC) program. The authors would also like to thank the Federal Ministry for Economic Affairs and Climate Action (BMWK) and the Project Management Juelich (PtJ) for funding the project "Increasing energy efficiency in production through digitization and AI" - ecoKI (03EN2047A).

References

1. Agrawal, A., Gans, J.S., Goldfarb, A.: What to expect from artificial intelligence. In: MITSloan Management Review Spring Issue (2017). https://sloanreview.mit.edu/article/what-to-expect-from-artificial-intelligence/
2. Alvela Nieto, M.T., Nabati, E.G., Bode, D., Redecker, M.A., Decker, A., Thoben, K.-D.: Enabling energy efficiency in manufacturing environments through deep learning approaches: lessons learned. In: Ameri, F., Stecke, K.E., von Cieminski, G., Kiritsis, D. (eds.) APMS 2019. IAICT, vol. 567, pp. 567–574. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29996-5_65
3. Heymann, H., Kies, A.D., Frye, M., Schmitt, R.H., Boza, A.: Guideline for deployment of machine learning models for predictive quality in production. Proc. CIRP **107**, 815–820 (2022). https://doi.org/10.1016/j.procir.2022.05.068
4. Huber, S., Wiemer, H., Schneider, D., Ihlenfeldt, S.: DMME: data mining methodology for engineering applications - a holistic extension to the CRISP-DM model. Proc. CIRP **79**, 403–408 (2018). https://doi.org/10.1016/j.procir.2019.02.106
5. MacDonald-Phillips, K.A., et al.: Communication skills training and assessment of food animal production medicine veterinarians: a component of a voluntary Johne's disease control program. J. Dairy Sci. **105**, 2487–2498 (2022). https://doi.org/10.3168/jds.2021-20677
6. Mayr, A., et al.: Machine learning in production - potentials, challenges and exemplary applications. Proc. CIRP **86**, 49–54 (2019). https://doi.org/10.1016/j.procir.2020.01.035
7. Potdar, A.M., Narayan, D.G., Kengond, S., Mulla, M.M.: Performance evaluation of docker container and virtual machine. Proc. Comput. Sci. **171**, 14191428 (2020). https://doi.org/10.1016/j.procs.2020.04.152
8. Rausch, T., Rashed, A., Dustdar, S.: Optimized container scheduling for data-intensive serverless edge computing. Future Gener. Comput. Syst. **114**, 259–271 (2021). https://doi.org/10.1016/j.future.2020.07.017
9. Salari M., Lina Kattan, L., William H.K. Lam, W.H.K., Mohammad Ansari Esfeh, M.A., Fu, H.: Modeling the effect of sensor failure on the location of counting sensors for origin-destination (OD) estimation. Trans. Res. Part C: Emerg. Technol. **132**(11), 103367 (2021). https://doi.org/10.1016/j.trc.2021.103367

10. Schäfer, F., Mayr, A., Schwulera, E., Franke, J.: Smart use case picking with DUCAR: a hands-on approach for a successful integration of machine learning in production processes. Proc. Manuf. **51**(1), 1311–1318 (2020). https://doi.org/10.1016/j.promfg.2020.10.183
11. Tang, S., Yang, Y.: Why neural networks apply to scientific computing? Theor. Appli. Mech. Lett. **11**(3), 100242 (2021). https://doi.org/10.1016/j.taml.2021.100242
12. Tuli, S., Tuli, S., Tuli, R., Gill, S.S.: Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. Internet Things **11**(10222), 567–574 (2020). https://doi.org/10.1016/j.iot.2020.100222
13. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. J. Big Data **3**, 9 (2016). https://doi.org/10.1186/s40537-016-0043-6

Autonomous Driving Based on Imitation and Active Inference

Sheida Nozari[1,2(✉)], Ali Krayani[1], Pablo Marin[2], Lucio Marcenaro[1], David Martin[2], and Carlo Regazzoni[1]

[1] University of Genoa, Genoa, Italy
sheida.nozari@edu.unige.it
[2] Carlos III University of Madrid, Getafe, Spain

Abstract. We advance a novel computational model of acquiring a hierarchical action sequence and its use for minimizing the divergence between observation and prediction. The model is grounded in a principled framework to learn and cognize the dynamic surrounding of a learning agent, which exemplifies the model's functioning by presenting a simulation of overtaking scenarios in autonomous driving. The learning agent integrates imitation learning and active inference to form hierarchical representations from expert demonstrations. During the online learning phase, the learning agent improves the action selection procedure based on the prior knowledge (exploiting) and the novel interactions with a dynamic environment (exploring). The proposed method applies an active knowledge sampling during the learning agent's movements to make a dynamic inference in the Bayesian structure by message passing through the multi-levels. A line-changing driving scenario with different levels of complexity is organized to verify the proposed framework's efficiency by dealing with single and multiple dynamic objects interacting in the environment.

Keywords: Imitation Learning · Motion prediction · Active inference · Dynamic Bayesian Network

1 Introduction

As autonomous vehicle (AV) technology progresses towards full autonomy, self-driving vehicles are more likely to become a reality in the coming years [1]. Although AVs have demonstrated a guarantee to operate in a structured environment under reasonable driving circumstances, they are still facing unpredictable behaviours causing disastrous consequences in unseen environments with high uncertainty [2]. A crucial component of autonomous driving is making driving decisions based on the vehicle's environmental changes and adapting to different types of driving conditions [3].

Due to their heavy manual tuning in modelling dynamic interactions among vehicles, traditional planning methods are not always scalable and cost-effective

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 13–22, 2023.
https://doi.org/10.1007/978-3-031-16281-7_2

[4]. Imitation Learning (IL) offers an alternative solution that mimics an expert's driving behaviour based on a data-driven approach. However, in IL, one of the primary challenges is specifying how an AV acts when it is outside the presented states by the demonstrations [5]. Reinforcement Learning (RL) can overcome such an issue by exploring new states and actions while an agent is interacting with the environment. Exploration plays a vital role in RL, however, it typically requires a large amount of training data and computational resources to succeed [6]. In response to this challenge, IL leverages inverse reinforcement learning (IRL) to deduce the reward function from demonstrations [7]–[8]. Learning a parameterized reward function provides a compact representation of the demonstrator's preferences and allows policy optimization to generalize to the unnoticed states. However, IRL methods always generate uncertainty about the proper reward function, which can have adverse outcomes if the learning agent figures a reward function leading it to learn the wrong policy. Additionally, systems that uses deterministic models rather than probabilistic ones ignore the stochastic property of dynamic environments. We propose an IL approach combined with a hierarchical probabilistic model under the Active Inference framework [9] to overcome the mentioned limitations. An active imitator agent (i.e., an AV) aims to learn a robust policy to uncertainty and dynamic environmental changes while accomplishing a specific task. The agent can effectively trade-off the exploration-exploitation rate with the expected return within active inference. The latter allows an AV to detect novel situations by comparing what the agent is expecting to observe based on the rules learned from an expert's demonstration and what it is actually observing to decide whether to explore new actions or exploit what has been learned by imitating the expert. AV aims to learn a sequence of actions that lead to the minimization of the abnormality. The main contributions of this work are: *1)* Integrate Active inference to IL to optimize the learning policy by establishing the balance between expected return and abnormality value under the new experiences. *2)* Optimize the exploration-exploitation rate by distinguishing the normal and abnormal situations through the Bayesian predictive and diagnostic messages. *3)* Optimize the action planning concerning the uncertainty and minimize the imitation loss with respect to expert demonstrations. *4)* Demonstrate that the proposed framework achieves better results in terms of learning aspects than existing RL methods.

2 Proposed Framework

The proposed framework is divided into two main phases: offline and online learning. During the first phase, we learn a situation model explaining how an *expert agent* (E) and a *dynamic Object* (O) interact in the environment. Moreover, we organize a First-Person generative model (FP), allowing a *learning agent* (L) to learn the E's behavior by observing its demonstrations. In the second phase, L modifies and updates its imperfect knowledge from the sub-optimal expert demonstration through an Active First-Person generative model (AFP) with respect to its relative distance with a moving object ($\hat{\text{O}}$) in a dynamic environ-

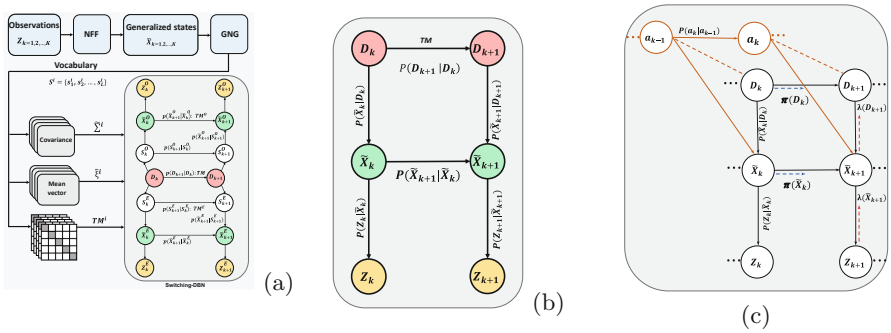


Fig. 1. Learning models: a) Situation model, b) FP model, c) AFP model.

ment. Both FP and AFP models are structured in Dynamic Bayesian Network (DBN) representations [10].

2.1 Offline Learning Phase

Situation Model. This model is structured to explain the interaction among two dynamic agents (see Fig. 1-(a)), E and O by using a switching DBN [11]. In order, the variables $\mathrm{Z}_k^{\mathrm{E}}$ and $\mathrm{Z}_k^{\mathrm{O}}$ illustrate the agents observations at the low level of the hierarchy (yellow nodes). The middle level (green nodes), describes the joint hidden continuous Generalized states (GSs) containing the dynamics of two agents at each time instant k as follows: $\tilde{\mathbf{X}}_k = [\tilde{\mathrm{X}}_k^{\mathrm{E}}\ \tilde{\mathrm{X}}_k^{\mathrm{O}}]^\intercal$, where $\tilde{\mathrm{X}}_k^{\mathrm{E}}$, $\tilde{\mathrm{X}}_k^{\mathrm{O}}$ denote the GSs of E and O, respectively. The GS related to agent i is defined as a vector composed of the agent's state and its first-order temporal derivative, such that $\tilde{\mathrm{X}}_k^i = [\boldsymbol{x}\ \dot{\boldsymbol{x}}]^\intercal$ where $\boldsymbol{x} \in \mathbb{R}^d$, $\dot{\boldsymbol{x}} \in \mathbb{R}^d$, $i \in \{\mathrm{E}, \mathrm{O}\}$ and d stands for the dimensionality. The correlation between Z_k^i and $\tilde{\mathrm{X}}_k^i$, which describes the observation model, is defined as:

$$\mathrm{Z}_k^i = \mathrm{H}\tilde{\mathrm{X}}_k^i + v_k, \tag{1}$$

where $H = [\mathrm{I}_d\ \ 0_{d,d}]$ stands for the observation matrix that measures of how dependent the measurements (Z_k^i) are upon the hidden GSs ($\tilde{\mathrm{X}}_k^i$) and $v_k \sim \mathcal{N}(0, \mathrm{R})$ is the measurement noise that follows a zero-mean Gaussian distribution with covariance R.

Initially, the evolution of $\tilde{\mathbf{X}}_k$ is assumed to follow a static equilibrium assumption described by:

$$\tilde{\mathrm{X}}_k^i = \mathrm{A}\tilde{\mathrm{X}}_{k-1}^i + w_k, \tag{2}$$

where $\mathrm{A} \in \mathbb{R}^{d \times d}$ and $w_k \sim \mathcal{N}(0, \mathrm{Q})$ depict the dynamic matrix and the process noise, respectively. $\tilde{\mathbf{X}}_k$ is predicted by utilizing a Null Force Filter (NFF) according to (2). The NFF calculates the innovations that encode the deviations between predictions and observations as : $\varepsilon_{\tilde{\mathrm{X}}_t} = \mathrm{H}^{-1}\big(\mathrm{Z}_k^i - \mathrm{H}\tilde{\mathrm{X}}_k^i\big)$. The Growing Neural Gas with utility measurement (GNG-U) [12] is empolyed to cluster

in an unsupervised manner those innovations $\varepsilon_{\tilde{\mathrm{X}}_t}$ that characterise the generalized errors (GEs) and outputs a vocabulary defined as: $\mathbf{S}^i = \{\mathrm{s}_1^i, \mathrm{s}_2^i, \ldots, \mathrm{s}_{\mathcal{L}_i}^i\}$ consisting of $\mathcal{L}_i$ clusters.

The interaction between E and O at multi-level (i.e., discrete and continuous levels) is described by a joint vocabulary (i.e., $\mathbf{S}^{\mathrm{E}}, \mathbf{S}^{\mathrm{O}}$) expressing discrete regions with quasi-linear models that explain the interactive dynamic evolution of joint states over time. Each discrete cluster $\mathrm{s}^i \in \mathbf{S}^i$ follows a multivariate Gaussian distribution with covariance matrix $\tilde{\Sigma}_{\mathrm{s}_k^i}$ and generalized mean value $\tilde{\boldsymbol{\mu}}^{\mathrm{s}^i} = [\mu_{Pos}^{s^i}\ \mu_V^{s^i}]$, where $\mu_{Pos}^{s^i}$ is the states' mean value on positions and $\mu_V^{s^i}$ is the states' mean value on velocity. Thus, a dictionary can be formed as $\mathbf{D} = \{\mathrm{D}_1, \mathrm{D}_2, \cdots, \mathrm{D}_{\mathrm{M}}\}$ according to the learned configurations, where the interaction configuration ($\mathrm{D}_k = [\mathrm{s}_k^{\mathrm{E}}, \mathrm{s}_k^{\mathrm{O}}]^\intercal$) explains the joint activated clusters.

Consequently, an extra vocabulary encoding the joint configurations is defined as: $\mathbf{D} = \{\mathrm{D}_1, \mathrm{D}_2, \cdots, \mathrm{D}_{\mathrm{M}}\}$ with M configurations, where $\mathrm{D}_k \in \mathbf{D}$ and $\mathrm{D}_k = [\mathrm{s}_k^{\mathrm{E}}, \mathrm{s}_k^{\mathrm{O}}]^\intercal$ depicts an interaction configuration explaining the joint activated clusters occurring simultaneously in the agents' vocabularies (red nodes). Each configuration D_k consists of the average position and average velocity of the two agents according to: $\mathrm{D}_k = \left[(\mu_{Pos}, \mu_V)^{\mathrm{E}}, (\mu_{Pos}, \mu_V)^{\mathrm{O}}\right]$. After learning the joint vocabulary, the dynamic model defined in (2) can be updated as follows:

$$\tilde{\mathrm{X}}_k^i = \mathrm{A}\tilde{\mathrm{X}}_{k-1}^i + \mathrm{B}\mu_V^{s_k^i} + w_k, \tag{3}$$

where $\mathrm{B} \in \mathbb{R}^{d \times d}$ stands for the control model matrix, $\mu_V^{\mathrm{s}_k^i} = [\dot{x}_{s_k^i}, \dot{y}_{s_k^i}]$ is a control vector encoding the agent's velocity (on x and y) associated with s_k^i. The Transition Matrix (TM) can be learned by estimating the transition probabilities $\mathrm{P}(\mathrm{D}_{k+1}|\mathrm{D}_k)$ that encodes the dynamic transitions among the learned configurations at the top level of hierarchy is defined as:

$$\mathrm{TM} = \begin{bmatrix} \mathrm{P}(\mathrm{D}_1|\mathrm{D}_1), & \mathrm{P}(\mathrm{D}_1|\mathrm{D}_2), & \ldots, & \mathrm{P}(\mathrm{D}_1|\mathrm{D}_{\mathrm{M}}) \\ \mathrm{P}(\mathrm{D}_2|\mathrm{D}_1), & \mathrm{P}(\mathrm{D}_2|\mathrm{D}_2), & \ldots, & \mathrm{P}(\mathrm{D}_2|\mathrm{D}_{\mathrm{M}}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{P}(\mathrm{D}_{\mathrm{M}}|\mathrm{D}_1), & \mathrm{P}(\mathrm{D}_{\mathrm{M}}|\mathrm{D}_2), & \ldots, & \mathrm{P}(\mathrm{D}_{\mathrm{M}}|\mathrm{D}_{\mathrm{M}}) \end{bmatrix} \tag{4}$$

where $\sum_m^{\mathrm{M}} \mathrm{P}(\mathrm{D}_p|\mathrm{D}_m) = 1$ such that $p, m \in \mathrm{M}$.

First-Person Model. We provide the FP model by using a switching DBN (situation model) represented by a generative DBN to consider the interaction states with a dynamic object ($\hat{\mathrm{O}}$) under the L's interpretation (see Fig. 1-(b)). The discrete level represents the learned configuration $\mathrm{D}_m \in \mathbf{D}$ (red nodes). The continuous level stands for the generalized relative distance between E and O, which will be updated by L and $\hat{\mathrm{O}}$ through interacting in the environment(green nodes). At FP model initialization, the generalized relative distance can be computed by the difference of the joint GS to consider the agent's interaction at a specific configuration at time k, as follow: $\tilde{\mathrm{X}}_k = \left[\tilde{\mathrm{X}}_k^{\mathrm{E}} - \tilde{\mathrm{X}}_k^{\mathrm{O}}\right] = \left[(\boldsymbol{x}^{\mathrm{O}} - \boldsymbol{x}^{\mathrm{E}})(\dot{\boldsymbol{x}}^{\mathrm{O}} - \dot{\boldsymbol{x}}^{\mathrm{E}})\right]$.

At the bottom level (yellow nodes), the observation can be mapped onto observations (Z_k^i) of both agents as: $\mathrm{Z}_k = \left[\mathrm{Z}_k^{\mathrm{E}} - \mathrm{Z}_k^{\mathrm{O}}\right]$.

2.2 Online Learning Phase

In this section, we aim to enhance IL's efficiency through a probabilistic hierarchical model in a dynamic environment. Therefore, the FP model must compound with the active states to minimise the divergence between the internal predictions and observations while acting in the environment.

Active First-Person Model. The AFP model takes as input a sequence of observed or previously learned configurations ($D_{1:M}$), an action based on the current configuration (a_k), and a random variable (Z_k) which represents the learner observation by its exteroceptive sensor at time instant k. The observation provides a relative distance between L and $\hat{\mathrm{O}}$ that is embedded in the continuous level ($\tilde{X}_k$) of the online learning model. L sustains an internal dynamic representation $\mathrm{P}(\mathrm{Z}, \tilde{\mathrm{X}}, \mathrm{D}, a)$ of the external environment encoded in an AFP model (see Fig. 1-(c)) and purposes to implicitly reduce the mismatch between what it is expecting to receive from the environment and what it is actually perceiving.

Action Selection Probability. At each time instant, L evaluates its current situation. In one condition, L has no prior knowledge about the current interaction with $\hat{\mathrm{O}}$. In the other condition, L has been informed that E was in the same situation while interacting with O (prior belief). Thus, L estimates the proper behavior whether to explore by performing new actions based on the current configuration (in the first condition) or exploit from the learnt configuration (in the second condition). The divergence between the observation and prediction is the main criteria which guides the action selection procedure. L employs Particle Filter (PF) to predict the learnt configuration with the least divergence until time k. Likelihood messages ($\lambda(\tilde{\mathrm{X}}_k)$ and $\lambda(\mathrm{D}_k)$) passing in a backward manner from the bottom level towards higher levels inside the AFP enable measuring the anomaly between the prediction of the propagated particles and the learner's observation, which is computed by cosine similarity ($\cos(\theta)$), as follow:

$$\cos(\theta) = \frac{\tilde{\mathrm{Z}}_k \ . \ \tilde{\mathrm{X}}_{k,n}}{||\tilde{\mathrm{Z}}_k|| \ ||\tilde{\mathrm{X}}_{k,n}||}. \tag{5}$$

The computed abnormality assesses the similarity of the current observation with predictions, and the corresponding configuration to the particle with the highest weight (the least anomaly) is assigned as the activated configuration ($\mathring{\mathrm{D}}$). The particles' weights can be updated by using $\lambda(\mathrm{D}_k)$ which is defined as: $\lambda(\mathrm{D}_k) = \lambda(\tilde{\mathrm{X}}_k)\mathrm{P}(\tilde{\mathrm{X}}_k|\mathrm{D}_k)$, where $\lambda(\tilde{\mathrm{X}}_k) = \mathrm{P}(\mathrm{Z}_k|\tilde{\mathrm{X}}_{k,n})$ is a multivariate Gaussian distribution such that $\lambda(\tilde{\mathrm{X}}_k) \sim \mathcal{N}(\mathrm{Z}_k, v_k)$ and $\lambda(\mathrm{D}_k)$ is a discrete probability distribution. Additionally, L considers two parameters to perform an action. One is the exploration rate (ϵ) where the highest particle weight (α) presents a control input on it as follows: $\epsilon_k = 1 - \alpha_k$. Both ϵ_k and α values are in an interval between 0 and 1, as much as α tends to 1 (higher similarity), ϵ_k goes to

0 (less exploration). L learns to decrease the exploration rate while minimizing the distinction between the prediction and observation. The other parameter is a threshold (ρ) which is obtained by a trial-and-error process. L compares ρ and ϵ_k to select an action at time k as below:

$$a_k \sim \begin{cases} \arg\max_{a_k} \mathrm{Q}(\mathcal{A}, \mathring{D}_k), \text{ if } \epsilon < \rho \ \ \text{(exploitation)}, \\ \textit{random} \ \ \text{from } \mathcal{A}^+, \text{ if } \epsilon \geq \rho \ \text{(exploration)}, \end{cases} \tag{6}$$

where $\mathcal{A} = \{\mathcal{A}^{\mathrm{E}}, \mathcal{A}^+\}$, as $\mathcal{A}^{\mathrm{E}} = \{a_1^{\mathrm{E}}, a_2^{\mathrm{E}}, \ldots, a_{\mathrm{Y}}^{\mathrm{E}}\}$ is a set of actions performed by L and encoded in the situation model that L aims to imitate during exploitation and $\mathcal{A}^+ = \{a_1, a_2, \ldots, a_8\}$ is a set of actions realizing 8 cardinal and ordinal directions which L selects during exploration. Moreover, during exploration, L records the new experiencing pair ($\mathrm{D}_k^+, a_k^+ \in \mathcal{A}^+$) in the Q-table incrementally.

Free Energy Measurement. The AFP model by leveraging a hierarchical structure computes the imitation cost using bottom-up (λ) and lateral messages (predictions by inter-slice links π) that drive posterior agent movements toward a better prediction to optimize the Free Energy (FE), which will be updated after gathering novel observations. During the online phase, L learns an optimized mapping between the Bayesian message passing, which causes a sequence of observations with a minimum distinction between the expectation and likelihood. The distinction can be estimated between the predictive message $\pi(\tilde{\mathrm{X}}_k)$ and the diagnostic message $\lambda(\tilde{\mathrm{X}}_k)$ after performing an action (a_{k-1}). We employ Kullback Leibler-Divergence ($\mathcal{D}_{\mathcal{KL}}$) between $\pi(\tilde{\mathrm{X}}_k)$ and $\lambda(\tilde{\mathrm{X}}_k)$ to calculate the FE after each performed action, as:

$$\mathcal{F} = \mathcal{D}_{\mathcal{KL}}\Big(\lambda(\tilde{\mathrm{X}}_k)||\pi(\tilde{\mathrm{X}}_k)\Big) = \int \lambda(\tilde{\mathrm{X}}_k) \log\left(\frac{\lambda(\tilde{\mathrm{X}}_k)}{\pi(\tilde{\mathrm{X}}_k)}\right) \mathrm{d}\tilde{\mathrm{X}}_k. \tag{7}$$

Action Selection Probability Update. At each time instant k, the model evaluates the performed action (a_{k-1}) with respect to the FE measurement and updates the selection probabilities in the probabilistic Q-table according the following equation:

$$\mathrm{Q}_k^* = (1-\eta)\mathrm{P}(a_{k-1}|\mathrm{D}_{k-1}) + \eta\Big[(1-\mathcal{F}_k) + \gamma \max_{a_k} \mathrm{P}(a_k|D_k)\Big], \tag{8}$$

where η is the learning rate which controls to what extent the new experiences overrides the previously recorded situations $(1-\mathcal{F}_k)$ is the normalized reward measurement with a range in $[0, 1]$, and γ is a discount factor. Our objective is to minimize the long term loss by keeping down the $\mathcal{F}$ though improving the action selection procedure.

3 Experimental Evaluation

Our framework is validated using two autonomous vehicles' multisensory information, 'iCab 1' and 'iCab 2' [13]. To consider the lane-changing scenario, the odometry module obtains positional information and velocity from the vehicles, where iCab 2 overtakes iCab 1 from the left side without colliding.

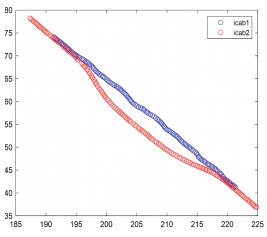
3.1 Offline Learning Phase

This section shows how the situation model is structured by employing NFF as an initial filter on data. The provided GEs by NFF are clustered using GNG that outputs a set of discrete clusters representing the discrete regions of the trajectories generated by E and O. The joint clusters introduce a set of configurations that encode the interactive behavior between the agents. (see Fig. 2).

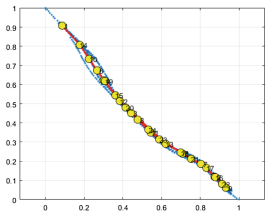
3.2 Online Learning Phase

The learning agent employs the FP model as a prior beliefs to imitate the expected transitions. During the online phase, by balancing exploration-exploitation trade-off, L's actions are engaged to solve the uncertain aspects of the action selection caused by the new dynamic environment, which adjust the L's hypotheses. The experiments are executed in a simulated environment through 500 episodes with different start positions to train a learning-agent L. Each episode consists of 10 iterations, i.e., L tries 5k iterations by 500 different start positions to learn the policies. We evaluate the performance of the proposed framework and compare it with other learning algorithms from the literature, namely, the general Q-learning, IRL (when an optimal expert is available), and self-learning in the RL context (when optimal expert data is not available).

Performance Evaluation. After trial stage, L acquires knowledge about the contingencies and the likelihood mapping in the generative model is aligned adequate with the reference generative process and the targeted goal (e.g., overtaking the dynamic object). Crucially, we assume that the correctness and accuracy of the action selection procedure guide the learning agent to the expected observations. Figure 3-(a) illustrates that L movements are engaged coherently, which causes less exploration in each trial epoch (e.g., each episode). Additionally, Fig. 3-(a) compares the number of executed actions during the training using different learning methods, where it shows L performs less actions to accomplish its task by using our method than others. Moreover, L adopting the proposed method has higher successful trajectories than other methods as depicted in Fig. 3-(b).



(a)



(b)

Fig. 2. Learning the situation model. a) iCab2 overtakes iCab1, b) Clustering of GEs.

Another noticeable point is that the AFP model expands its repertoire of action representations by learning new interactions between L and $\hat{O}$. The novel experienced configurations are recorded in Q-table incrementally, which will be clustered by using GNG after training. Figure 4-(a)-(b) shows the learned model during online phase has an expanded transition matrix than the situation model due to the exploratory aspect of the L's behavior.

Learning Cost Evaluation. The loss cost reduction imperative is one of the components of updating the actions' probabilities (see (8)) that guides the action selection policies in active inference. Figure 5-(a) demonstrates how modifying the actions can reduce the exploration and minimize the imitation cost resulting in a high learning rate during the training phase. Our goal is to find the best set of actions that minimize the imitation loss in terms of FE. Figure 5-(b) shows that the normalized global FE ($\mathcal{F}$) drops down capably to decrease below 0.1. Figure 5-(c) demonstrates that our method outperforms others in terms of suc-

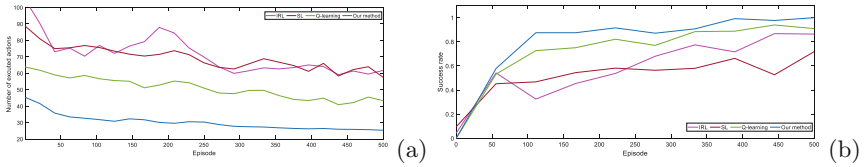


Fig. 3. a) The performed actions by L, b) The success rate.

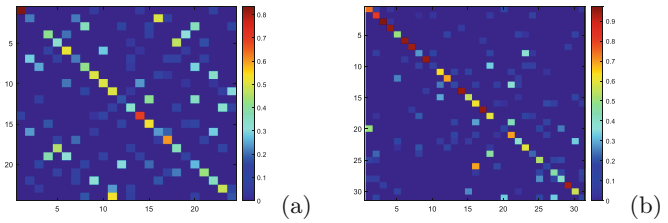


Fig. 4. Incremental learning. a) Corresponding TM to the situation model, b) TM after clustering the learnt Q-table.

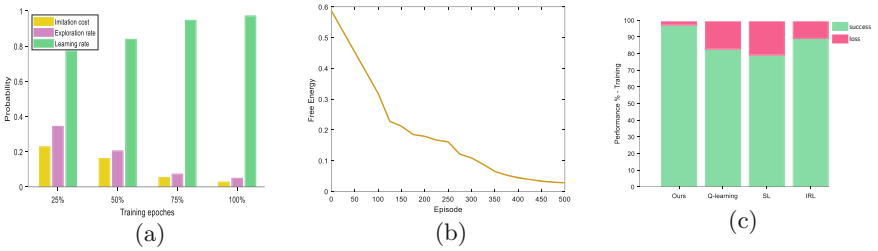


Fig. 5. Learning evaluation. a) The impact of exploration and learning rates after each training quarter on the imitation cost, b) FE measurement, and c) Training result.

cess and loss rate, which is attributed to the effectiveness of motion prediction while dealing with abnormalities that improve the success rate.

During testing, the agent travels through 500 paths with different start positions than the training time. The testing stage includes two levels of difficulties: I) the agent needs to overtake a single dynamic object, and II) the agent needs to overtake multiple dynamic objects. Moreover, each level has three scenarios: overtaking from the left side, overtaking from the right side, and when the agent should decide to overtake from the left or right side of the object(s). Table 1 describes that the adapted agent to the learned model through the presented method can overtake a single object and multiple dynamic objects in the environment effectively, whereas other methods still have a high failure rate. Moreover, the provided vocabulary by E's demonstrations tends to change-line from the left side of O. By experiencing the unseen configuration during the online learning phase, L learns to reduce the collision probability by interacting from the dynamic object's right side. The testing results show that by $5k$ training trajectories, the agent can overcome the experiments with different scenarios where it is necessary i) to overtake from the left side of the dynamic object(s), ii) to overtake from the right side of the dynamic object(s), and iii) to decide to overtake from which side of the dynamic object(s) has less collision probability (mixed situation from both side).

Table 1. Testing the learnt model after $5k$ trial trajectories.

	Our method		Q-learning		IRL		SL	
	Overtake%	Collision%	Overtake%	Collision%	Overtake%	Collision%	Overtake%	Collision%
	Difficulty I: overtaking from one dynamic object							
From the left side	**97.42**	**2.58**	88.13	11.87	90.28	9.72	79.67	20.33
From the right side	**96.81**	**3.19**	84.01	15.99	89.04	10.96	71.52	28.48
From both side	**94.86**	**5.14**	81.35	18.65	86.99	13.01	69.43	30.57
	Difficulty II: overtaking from two dynamic objects							
From the left side	**95.26**	**4.74**	83.78	16.22	89.11	10.89	73.51	26.49
From the right side	**93.13**	**6.87**	80.28	19.72	85.33	14.67	66.17	33.83
From both side	**92.05**	**7.95**	79.15	20.85	80.06	19.94	64.09	35.91

4 Conclusion

A novel framework has been proposed to integrate Imitation Learning with Active Inference for autonomous driving. In the hybrid presented model, the errors between the prediction and observation guide the action selection in two processes. A low amount of error influences L that exploits the prior knowledge to perform an action. In this case, L prepares an imitative response. On the other hand, when experiencing an unobserved configuration causes a high error, L needs to rely on random movements. During the online phase, L learns to decide how to minimize the FE and guide the movements to the imitative actions. Future work concentrates on employing the errors to guide the random movement that might facilitate the execution of congruent actions with the expert demonstrations.

References

1. Claussmann, L., Revilloud, M., Gruyer, D., Glaser, S.: A review of motion planning for highway autonomous driving. IEEE Trans. Intell. Transp. Syst. **21**(5), 1826–1848 (2020)
2. Guo, J., Kurup, U., Shah, M.: Is it safe to drive? an overview of factors, metrics, and datasets for driveability assessment in autonomous driving. IEEE Trans. Intell. Transp. Syst. **21**(8), 3135–3151 (2020)
3. Li, Z.: A hierarchical autonomous driving framework combining reinforcement learning and imitation learning. In: 2021 International Conference on Computer Engineering and Application (ICCEA), pp. 395–400, June 2021
4. Zhou, J.: Exploring imitation learning for autonomous driving with feedback synthesizer and differentiable rasterization. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1450–1457 Sep. 2021
5. Hussein, A., Gaber, M.M., Elyan, E., Jayne, C.: Imitation learning: a survey of learning methods. ACM Comput. Surv. (CSUR) **50**(2), 1–35 (2017)
6. Sutton, R.S. Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
7. Ng, A.Y., Russell, S.J., et al.: Algorithms for inverse reinforcement learning. In: Icml, vol. 1, p. 2 (2000)
8. Nozari, S., Marcenaro, L., Martin, D., Regazzoni, C.: Observational learning: imitation through an adaptive probabilistic approach. In: 2021 IEEE International Conference on Autonomous Systems (ICAS IEEE 2021), pp. 1–5
9. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G.: Active inference: a process theory. Neural Comput. **29**(1), 1–49 (2017)
10. Ghahramani, Z.: Learning dynamic Bayesian networks. In: Giles, C.L., Gori, M. (eds.) NN 1997. LNCS, vol. 1387, pp. 168–197. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0053999
11. Baydoun, M., Campo, D., Kanapram, D., Marcenaro, L., Regazzoni, C.S.: Prediction of multi-target dynamics using discrete descriptors: an interactive approach. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3342–3346. IEEE (2019)
12. Iqbal, H., Campo, D., Baydoun, M., Marcenaro, L., Gomez, D.M., Regazzoni, C.: Clustering optimization for abnormality detection in semi-autonomous systems. In: 1st International Workshop on Multimodal Understanding and Learning for Embodied Applications, pp. 33–41 (2019)
13. Marin-Plaza, P.: Stereo vision-based local occupancy grid map for autonomous navigation in ros. In: 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications VISIGRAPP, vol. 2016 (2016)

Machine Learning Based Reconstruction of Process Forces

Berend Denkena, Heinrich Klemme, and Dennis Stoppel(✉)

Institute of Production Engineering and Machine Tools, Leibniz Universität Hannover, An der Universität 2, 30823 Garbsen, Germany
stoppel@ifw.uni-hannover.de

Abstract. During milling, process forces are acting on the cutting tool, causing tool deflection and subsequently a shape deviation of the workpiece. To compensate these effects, knowledge of the process forces is required. In this work, machine learning (ML) methods are applied to reconstruct process forces from the drive signals of two different milling centers. The results of a linear regression, bagged trees and a stacked LSTM are presented. The approaches show different results depending on the milling center. Only for the LSTM an error lower than 30 N is achieved for both machine tools. Independent of the ML approach, the results strongly depend on the selection of milling processes used for training.

Keywords: Machine-learning · Milling · Artificial neural network · Machine tool

1 Introduction

In machining, the workpiece quality is significantly related to the process forces. E.g. process forces are decisive for the displacement of tool and workpiece [1, 2]. With a process force-based control, it is possible to minimize resulting shape deviation by adapting the process parameters [3, 4]. There exist various methods for measuring process forces. In general, the closer the force is measured to the tool center point (TCP), the lower the interference on the measured variable. A distinction is made between direct and indirect force measurement.

Direct methods are defined by measuring the force close to the cutting region [5]. They include the measuring principles of the piezoelectric effect using dynamometers and the strain-dependent change in resistance of strain gauges. Due to their high accuracy, dynamometers are often used in research [5]. For industrial applications, the dynamometer is rarely used due to high costs and workspace limitation [6]. They also affect the compliance of the tool-workpiece-machine-system and are susceptible to drift [7]. In addition to dynamometers, strain gauges are used for direct force measurement. For example, Rizal [8] designed a tool holder with a force sensing element located between the tool holders' shaft and its taper, equipped with 24 strain gauges. Due to the tool holders natural frequency it is only applicable for use at spindle speeds below 5,000 rpm. In addition, the working space is reduced. Denkena and Boujnah integrated the strain

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 23–32, 2023.
https://doi.org/10.1007/978-3-031-16281-7_3

gauges into a spindle slide [3]. The strain gauges are attached within small notches on the spindle slide to increase the sensitivity of the system. The developed method has a measuring accuracy of 25 N.

Indirect methods can be subdivided in sensor-based and sensorless approaches. An example for sensor-based approaches is to measure the displacement of the spindle shaft with either inductive [9] or capacitive displacement sensors [10]. The forces are calculated based on the stiffness of the spindle structure. The advantage is that it neither affects the compliance nor restricts the workspace. By compensating the structural dynamics between tooltip and spindle flange with a Kalman filter, forces are measured up to a frequency of 1,000 Hz. A general disadvantage of sensor-based approaches is the probability of system downtimes due to sensor failures.

A common option for sensorless approaches is the Kalman filter-based disturbance force observer [5, 11]. Altintas et al. [12] provide a fundamental work, where they reconstruct process forces based on the feed drive current for frequencies up to 100 Hz. Kim et al. [13] used the drive currents of the servo motors of the x- and y-axis of a milling machine and compared the results of a Kalman filter-based disturbance force observer to an artificial neural network (ANN). They found that processes with a reversal of direction, like circular pocket milling, are estimated more accurate by the ANN.

Schwenzer et al. [14] investigated approaches to correct the drive currents from friction, inertia and positions-dependent components. They compared model-based approaches based on a Stribeck model and a generalized Maxwell-slip friction model (GMS) to a support vector regression (SVR) and conclude that the SVR reaches the lowest root mean square error (RMS). Nevertheless, they emphasize that the selection of training data is of great importance to achieve a high generalizability.

In [15], a model-based approach is compared to a long short-term memory based neural network (LSTM). While the model-based approach shows good results for processes with constant feed velocity, the accuracy decreases significantly for circular axis movements. Comparing to the reference measurement, an increase in the error is observed when the feed direction reverses. The reason for this is the insufficient compensation of microfriction with the used Stribeck model, which occur especially at velocities close to zero. Dynamic friction models, as the LuGre model [9], provide a more accurate result. Still, influences like load and temperature are not included. Furthermore, position-dependent effects such as guide rail wear are not represented. Therefore, ML represent an important alternative to model-based approaches [13, 15]. In each case, the ANN learn from initial dynamometer measurements. A decisive point using ML is generalizability. The trained network should be able to transfer the learned force reconstruction to any other process. However, the effect of process variations on the prediction of ML models has not been investigated.

Therefore, three ML methods of different complexity are analyzed in this paper. First, linear regression is chosen as a baseline. Second, bagged tree regression is analyzed. Last, a LSTM according to [15] is evaluated. The network is optimized and applied to another milling center. In the following, the experimental procedure and the data acquisition are described with regard to the generalizability. Furthermore, the methods are compared regarding their prediction error. Finally, the generalizability is shown by transferring the methods to another milling machine.

2 Methods

In this section, the experimental setup and the cutting investigations as well as the data pre-processing are described.

2.1 Experimental Setup

Numerous cutting investigations with steel (AISI 1045) and aluminum (A7075 T6) were carried out on the 5-axis milling center MILLTAP700, without coolant. Three, resp. Two different solid carbide end mills were used for each material (Table 1).

Table 1. Solid carbide end milling cutter, steel: no. 1–3; aluminum: no. 4 & 5

No.	Diameter d [mm]	Helix angle $\varepsilon_{1/2}$ [°]	Rake angle γ_0 [°]	Number of teeth n [-]	Tool length l [mm]
1	6	35/38	8	4	62
2	8	35/38	8	4	68
3	10	35/38	8	4	80
4	6	40	13	3	60
5	10	40	13	3	70

In order to increase the variance of the data, width of cut, spindle speed and feed velocity were varied. Moreover, different milling operations are included. A distinction is made between circular and rectangular pockets and straight and curved paths in different directions. To cover different depths of cut and avoid overfitting, different surface profiles were first milled into the workpiece. As shown in Fig. 1, the resulting workpiece has three levels h1, h2 and h3 with different heights. After this preparation, cutting investigations are conducted to record training and test dataset. Each element of the dataset is a time series with a unique combination of process parameters, tool and workpiece material. The different signals listed in Table 2. Subsequently, a single time series is referred to as a process p.

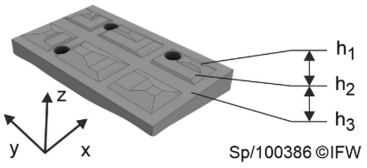


Fig. 1. Prepared workpiece

A Kistler dynamometer 9772B was used to record process forces during machining. The signal is amplified and analogue low-pass filtered by a Kistler laboratory charge

amplifier type 5015A with a cutoff frequency of 1 kHz. Thereby, the experimental setup is equal to [15] except for the milling center. Position, speed, acceleration and current of the x-, y- and z-axis were recorded by the machine control via SinuTrace. The sampling frequency is 250 Hz, limited by the interpolation cycle of the machine control.

2.2 Data Preprocessing

First, the recorded data is divided into training and test set. The subdivision was made such that a balanced distribution of the process parameters is present in both sets. To evaluate generalizability of the methods regarding to the workpiece material, cutting investigations with aluminum were only included in the test data such that the models are exclusively trained with steel as a workpiece material. The training dataset contains 161 and the test data set 25 different processes. The ML models input variables $x_i^p(k)$, with the index p of the respective measurement and the index i of the respective control signal, are listed in Table 2. Each measurement is recorded over a discrete period of time with the sample index k.

Table 2. Signal indices

Index (x / y / z)	Signal	Index (x / y / z)	Signal
i = 1 / 5 / 9	Position	i = 4 / 8 / 12	Current
i = 2 / 6 / 10	Velocity	j = 1 / 2 / 3	Force (dynamometer)
i = 3 / 7 / 11	Acceleration		

The corresponding measurement signal $y_j^p(k)$ of the dynamometer is used as output variable. The index j denotes the process force direction in the machine coordinate system. Since the input signals x_i^p have different value ranges and units, the inputs are normalized with the Z-score normalization that is widely used for ML methods (Eq. 1 [16]). For each drive signal variable mean value μ_i and standard deviation σ_i are calculated along all measurements of the training data set. The forces are not normalized. Output normalization was tested during hyperparameter optimization, but did not improve the result of the training for any approach.

$$x_{i,z}^p(k) = \frac{x_i^p(k) - \mu_i}{\sigma_i} \tag{1}$$

Due to the Shannon theorem, the measured force can be reconstructed up to a sampling frequency half of the input signals frequency. Consequently, the measurement results of the dynamometer are filtered with a cutoff frequency of 100 Hz by a Butterworth low-pass filter of fifth order.

2.3 Algorithms and Model Training

MATLAB R2021b is used to evaluate the experimental data. For simplification, only forces in x- and y-directions are analyzed. The normalized drive signals $x_{i,z}^p(k)$ are used

as input values. Separate models are trained for the different directions. First, linear regression $g_{LR}(x(k))$ is used as a baseline, see Eq. 2. Thus, the outcome of the linear regression model is the sum of the input signals $x_i(k)$ multiplied by a signal specific weight w_i plus a bias b. In the equation i denotes the signal index and n is the total number of input variables which equals 12. According to Sect. 1, the real system is expected to show a non-linear behavior due to friction. Consequently, linear regression can be used to investigate how well the described system can be linearized and how high the residual error is. In addition, $g_{LR}(x)$ defines the minimal accuracy every model should reach, since it is the least complex approach in this paper.

$$g_{LR}(\boldsymbol{x}(k)) = b + \sum_{i=1}^{n} w_i x_i(k) \tag{2}$$

In addition, bootstrap aggregation regression trees $g_{BT}(x)$ with an ensemble of 50 trees according to [17] are used. An ensemble of 50 trees was used. For the training, k-fold cross-validation with k = 5 was applied.

An ANN $g_{RNN}(x)$ based on LSTM cells is used as the model with the highest complexity. Due to the recurrence, the LSTM is the only approach that takes the temporal context of the input into account. The network topology of the LSTM is validated and optimized using the training data. For the final investigations, a network with 7 layers (5 hidden layers) prevailed. The input layer has 12 input neurons. These are followed by two recurrent layers with 30 LSTM cells each. The next three layers are fully connected layers with 50 neurons and the leaky ReLU activation function [18].

The last layer is the output of the network which predicts the force for the respective time step. For regularization, dropout is used between the hidden layers during training. The whole neural network is depicted in Fig. 2.

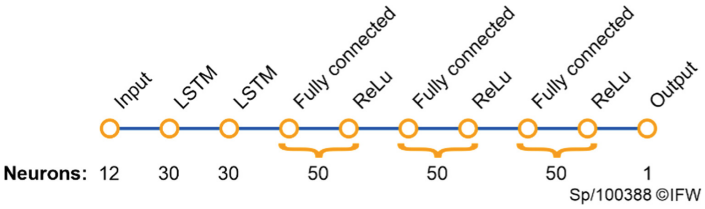


Fig. 2. Layers of the recurrent neural network

The network learns via backpropagation. The mean squared error (MSE) is used as error function, Eq. 3, to calculate the loss for the n_k discrete samples of each time series. Subsequently, an Adam optimizer is used to adapt the weights of the LSTM. The metrics used to evaluate the results are the RMSE and the mean absolute error.

$$MSE = \frac{\sum_{k=1}^{n_k} (y_i - g(x_i(k)))^2}{n_k} \tag{3}$$

3 Results

In the following, the results of the three different ML approaches with regard on the training data are presented and finally transferred to the HSC30 dataset from [15].

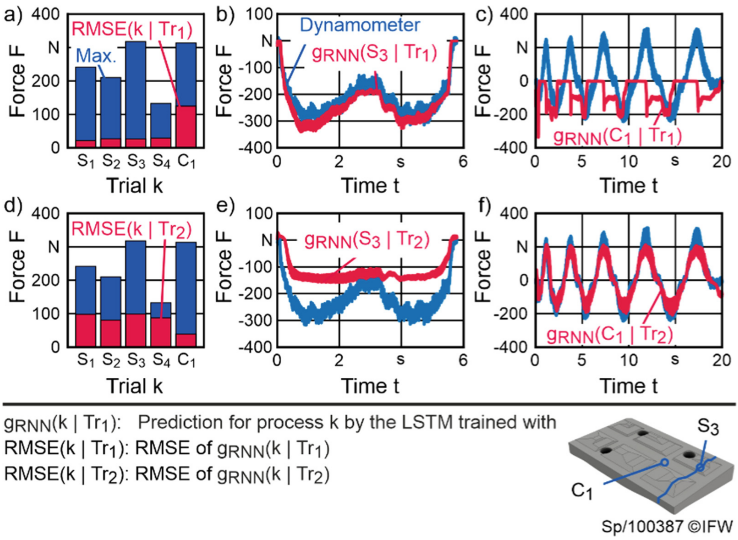


Fig. 3. Results of two different LSTM-networks trained with different milling processes. a) RMSE of the LSTM trained with notch milling (L1) b) notch milling S_3 compared to L1 c) circular pocket milling C_1 compared to L1 d) RMSE of the LSTM trained with circular pocket milling (L2) e) notch milling S_3 compared to L2 f) circular pocket milling C_1 compared to L2

3.1 Data Complexity

To analyze the impact of training data selection on the generalizability, a subset of the training data is examined. The subset is divided into four datasets: training dataset 1 (Tr_1), training dataset 2 (Tr_2), test dataset 1 (Te_1) and test dataset 2 (Te_2). Tr_1 and Te_1 contain only processes with constant feed velocity in x-direction, whereas Tr_2 and Te_2 contain circular pocket milling processes. Then, two identical LSTMs are trained, one with Tr_1 and another with Tr_2. Finally, both LSTMs are evaluated for Te_1 (processes S_1-S_4) and Te_2 (process C_1). For the LSTM trained with Tr_1 an error of 26.3 N is achieved for the force in x-direction, averaged over Te_1. The results are shown in detail in Fig. 3 a). However, the error increases significantly up to 125.0 N for Te_2. The analysis of the signal curves also reveals a high deviation between reconstructed and actual signal, Fig. 3 c). Most obvious, the LSTM is not capable of reconstructing positive forces, because it was not trained for this range. In addition, there is a change in the reconstructed force of C_1, when feed direction alters because the model was not taught with processes with slow velocities and alternating feed directions and is not able to learn e.g. the micro friction behavior of the respective axis. On the other hand, a model that is only trained with accelerated processes is not able to correctly reconstruct processes with constant feed velocity, as shown by the model trained with Tr_2, Fig. 3 d). Here, a significantly lower force is reconstructed, leading to a RMSE of 80 N to 100 N. The circular pocket milling is reconstructed with a lower RMSE of 39.3 N. As shown in Fig. 3 e), the reconstructed force is constantly smaller than the actual process forces. This indicates, that the training dataset did not contain enough data with forces greater than 200 N. In

addition, the dataset did not contain processes where acceleration is zero, which affects the learning of inertia dependent effects.

The results indicate, that a model can only be successfully generalized, if the training data represents all significant influences on the drive currents that result from the physical boundary conditions for the transmission behavior of the feed drives. The physical influences are significantly affected by the variation of axis position, speed and acceleration. In addition, there are load and temperature dependencies, which can vary. From the above, two rules can be derived for time series datasets:

1. The dataset has to be physically balanced. Physical effects that are not represented in the dataset or limited to rare incidents will not be learned by the ML approach.
2. Datasets should be statistically balanced. If the signals depend on reference values, variation of those values is crucial, to avoid extrapolation. The same is valid for the value range of the output signal.

3.2 Generalizability

The linear regression model shows the highest error during training and during evaluation using the test set. The RMSE of the x-axis is 63.3 N evaluating the training set and 60.5 N evaluating the test set. For the y-axis, a slightly higher accuracy is achieved, with an error of 63.3 N for training and 48.3 N for the test set, Fig. 4 a).

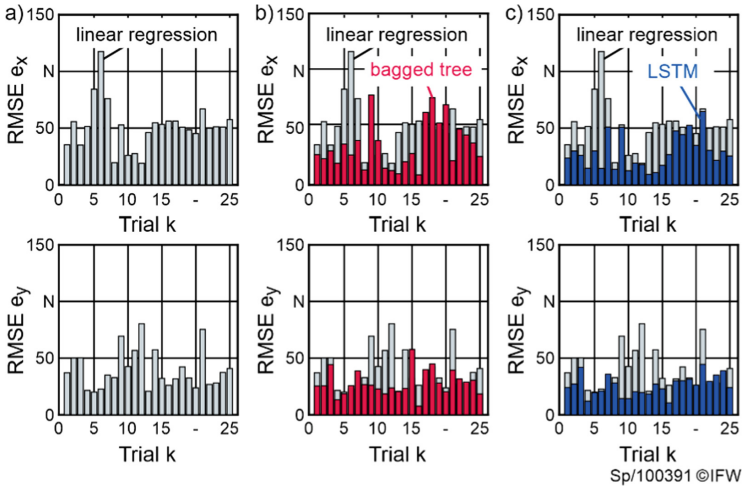


Fig. 4. RMSE of the test dataset a) linear regression b) bagged tree c) LSTM

The errors for the training and test sets are similar, indicating that the model is generalizable within the bounds of its accuracy. Nevertheless, the reconstruction has high errors. While there is no correlation of the RMSE to the process parameters, the linear model of the x-drive has a low accuracy for processes with constant feed velocity in

x-direction and a feed velocity of zero in y-direction. In contrast, the linear approximation of the y drive has a higher accuracy when reconstructing normal forces for the same section of processes. However, the error does not depend on the absolute force. Comparing the reconstructed force to the measurement, there are systematic deviations. When feed direction reverses and feed velocity is zero, there is a step in the regression curve, Fig. 5 a). Comparing the regressions output to the motor current, the current is decisive for the step. The current signals are heavily weighted for the linear model. In addition to the current, the feed velocity in the axial direction is weighted high, which can be interpreted as an approximation of viscous friction. For very low velocities and especially the reversal of feed direction, a linear model is insufficient. However, the error of the linear model shows a repetitive pattern depending on the velocity, Fig. 5 c).

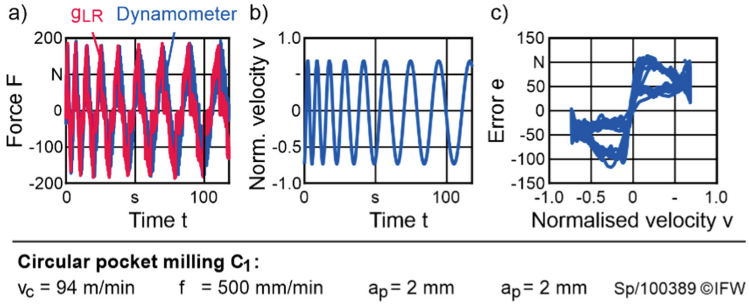


Fig. 5. Results of the linear regression for a pocket milling process a) Process forces b) Axis velocity c) Error between linear regression and dynamometer

Compared to the linear model, bagged trees achieve a 38% lower error, evaluating the test dataset, Fig. 4 b). For the training data, this results in an RMSE of 12.6 N for the reconstruction in the x-direction and an RMSE of 11.7 N in the y-direction. However, the errors are significantly higher for the test set with 41.9 N in the x-direction and 34.6 N in the y-direction. The large differences indicate overfitting. Consequently, the bagged tree must be regarded as not sufficiently generalizable.

The most accurate results are achieved with the LSTM, reaching a RMSE of 35.4 N in x- and 32.1 in y-direction (Fig. 4 c). All approaches were capable of reconstructing the forces for both workpiece materials, AISI 1045 (Fig. 4 c), trial 1–16) and A7075 T6 which was not used in the training dataset (Fig. 4 c), trial 17–25). On average, the process force is reconstructed with a RMSE of 22.6 N for AISI 1045, the RMSE is 35.5 N for A7075 T6. Consequently the RMSE is 36.2% lower for milling AISI 1045. Therefore a loss of accuracy cannot be ruled out.

Figure 6 a) depicts the comparison of all approaches. The accuracy increases with the complexity of the approach. Overall, the investigations show more accurate results for the y-axis. It is assumed that this is due to the machine tools' structure. Especially the ball screw shaft of the x-axis is much longer compared to the y-axis. As shown in Sect. 3.1, the training dataset has also an impact on the results of the learning algorithm, since the output signals used for the regression task differ.

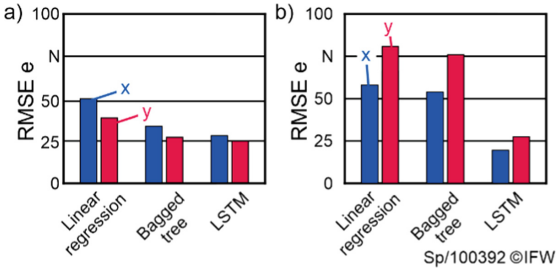


Fig. 6. RMSE for all three evaluated models a) MILLTAP700 b) HSC30

3.3 Comparing MILLTAP700 and HSC30

All in all the LSTM-based approach achieves the lowest RMSE. On average, the processes in the test dataset achieve a RMSE of 35.4 N for the x- and 32.1 N for the y-axis. When trained with the HSC30 dataset, the same architecture achieves a RMSE of 19.5 N in x- direction and 27.5 N in y-direction, which is 46.8% lower compared to [15]. The same preprocessing as proposed in Sect. 2 was applied. Due to the different amount of data samples, the number of training epochs was reduced in comparison to the training with the MILLTAP700 data. In general, the LSTM-based approach works for different machine tools with different drives, since the HSC30 is actuated by direct drives. This does not apply for the bagged tree and the linear regression, Fig. 6 b). The bagged tree achieves a RMSE of only 53.9 N in x-direction and 75.9 N in y-direction. The linear regression shows an even higher RMSE of 58.1 N in x- and 80.8 N in y-direction. It is assumed that the error of the linear regression and the bagged tree increases due to nonlinear cogging forces of direct drives.

4 Summary and Conclusion

It was shown that the LSTM-based force reconstruction developed in [15] can also be transferred to other machines. Even though, HSC30 and MILLTAP700 have different drives, the RMSE of the LSTM is similar. To train the LSTM, a dataset was recorded and the hyperparameters of the LSTM were optimized. The new LSTM was also trained with the HSC30 dataset. Compared to [15], a lower error was achieved. It is shown that the selection of the processes has a great importance for the training and the generalizability of the evaluated machine learning approaches. The training data set must have sufficient variation of the processes in the training data. Under this condition, transferability between workpiece materials from steel to aluminium could be demonstrated. On the other side, using data restricted to a specific type of process like circular pockets yields misleading results. It was shown that although apparently good results can be achieved with low variation of process parameters, these are not sufficiently transferable to other processes like unidirectional milling process.

Acknowledgements. We thank the “Sieglinde Vollmer Stiftung” for funding this research.

References

1. Brecher, C., Wetzel, A., Berners, T., Epple, A.: Increasing productivity of cutting processes by real-time compensation of tool deflection due to process forces. J. Mach. Eng. **19**, 16–27 (2019)
2. Wan, M., Zhang, W.H., Qin, G. H., Wang, Z.P.: Strategies for error prediction and error control in peripheral milling of thin-walled workpiece. Int. J. Mach. Tools Manuf. **48**(12–13), 1366–1374 (2008)
3. Denkena, B., Bergmann, B., Stoppel, D.: Tool deflection compensation by drive signal-based force reconstruction and process control. Procedia CIRP **104**, 571–575 (2021)
4. Denkena, B., Boujnah, H.: Feeling machines for online detection and compensation of tool deflection in milling. CIRP Ann. **67**(1), 423–426 (2018)
5. Aslan, D., Altintas, Y.: Prediction of cutting forces in five-axis milling using feed drive current measurements. IEEE/ASME Trans. Mechatron. **23**(2), 833–844 (2018)
6. Wan, M., Yin, W., Zhang, W.H.: Study on the correction of cutting force measurement with table dynamometer. Procedia CIRP **56**, 119–123 (2016)
7. Totis, G., Adams, O., Sortino, M., Veselovac, D., Klocke, F.: Development of an innovative plate dynamometer for advanced milling and drilling applications. Measurement **49**, 164–181 (2014)
8. Rizal, M., Ghani, J.A., Nuawi, M.Z., Haron, C.H.C.: Development and testing of an integrated rotating dynamometer on tool holder for milling process. Mech. Syst. Signal Process. **52**, 559–576 (2015)
9. Sarhan, A.A.D., Matsubara, A., Sugihara, M., Saraie, H., Ibaraki, S., Kakino, Y.: Monitoring method of cutting force by using additional spindle sensors. JSME Int J., Ser. C **49**(2), 307–315 (2006)
10. Albrecht, A., Park, S.S., Altintas, Y., Pritschow, G.: High frequency bandwidth cutting force measurement in milling using capacitance displacement sensors. Int. J. Mach. Tools Manuf **45**(9), 993–1008 (2005)
11. Yamato, S., Kakinuma, Y.: Precompensation of machine dynamics for cutting force estimation based on disturbance observer. CIRP Ann. **69**(1), 333–336 (2020)
12. Altintas, Y.: Prediction of cutting forces and tool breakage in milling from feed drive current measurements (1992)
13. Kim, T.Y., Woo, J., Shin, D., Kim, J.: Indirect cutting force measurement in multi-axis simultaneous NC milling processes. Int. J. Mach. Tools Manuf **39**(11), 1717–1731 (1999)
14. Schwenzer, M., Auerbach, T., Miura, K., Döbbeler, B., Bergs, T.: Support vector regression to correct motor current of machine tool drives. J. Intell. Manuf. **31**(3), 553–560 (2019). https://doi.org/10.1007/s10845-019-01464-1
15. Denkena, B., Bergmann, B., Stoppel, D.: Reconstruction of process forces in a five-axis milling center with a LSTM neural network in comparison to a model-based approach. J. Manufact. Mater. Process. **4**(3), 62 (2020)
16. Eesa, A.S., Arabo, W.K.: A normalization methods for backpropagation: a comparative study. Sci. J. Univ. Zakho **5**(4), 319–323 (2017)
17. Meinshausen, N., Ridgeway, G.: Quantile regression forests. J. Mach. Learn. Res. **7**(6) 983–999 (2006)
18. Maas, A.L., Hannun, A. Y., Ng, A.Y. Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of Icml, vol. 30 (1), pp. 3–9 (2012)

A Novel Rule-Based Modeling and Control Approach for the Optimization of Complex Water Distribution Networks

Enrico Ferrari[1(✉)], Damiano Verda[1], Nicolò Pinna[1], and Marco Muselli[1,2]

[1] Rulex Innovation Labs, 16122 Genoa, Italy
{enrico.ferrari,damiano.verda,nicolo.pinna}@rulex.ai
[2] Institute of Electronics, Computer and Telecommunication Engineering, National Research Council of Italy, 16149 Genoa, Italy
marco.muselli@ieiit.cnr.it

Abstract. This work applies Rule Based Control, a new rule-based, computationally efficient machine learning method for optimizing complex networks. This approach does not require a rigorous formulation of the optimization model, since only a set of historical data, where solutions are labeled as good or bad, is needed. It makes use of a rule-based machine learning method, which allows the optimization of complex networks, where a full description of the system is not available or too complex. The proposed approach is currently under evaluation in the water distribution system of the Milan (Italy) water main. The application of the approach to synthetic data shows its ability of reducing the energy consumption, while ensuring a good quality of service.

Keywords: Rule Based Control · Water distribution network · Classification · Optimization

1 Introduction

In the present work, a rule-based machine learning method, referred to as *Rule Based Control* (*RBC*), is proposed for real-time optimization in a water distribution network. The optimization of the whole network depends on the overall energy consumption and on the system quality of service. It is very difficult for a human operator to know in advance whether or not configuration changes made to a pump will result in improved performance, both in terms of energy consumption and quality of service. The quality of the service definition could be based on the quality of the water (e.g. chlorine concentration or water age) or it could be based on pressure management, that is on the compliance with the desired pressure level in each node and level of each tank.

Even if the Rule Based Control method is generic, and can be applied to any real time optimization problem, the present work is focused on optimizing the quality of service in terms of pressure management, whose relevance for water

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 33–42, 2023.
https://doi.org/10.1007/978-3-031-16281-7_4

main operations on a daily basis is discussed in Ghorbanian et al. [2016]. The proposed approach involves two phases (a) building a predictive model described by intelligible rules for the network status (e.g. *not fulfilling constraints*, *consuming too much energy* and *efficient use of energy*) (b) changing the status of the pumps trying to *validate* rules that predict the target condition(s).

Rule Based Control is currently used in a production context by the water distribution system of the Milan water main, which is evaluating the performances in terms of energy savings and pressure constraints fulfillment in comparison with state-of-the-art solutions. The paper is organized as follows: Sect. 2 discusses related work in this field, Sect. 3.1 describes the adopted modeling technique and 3.2 shows how this enables the development of an innovative rule-based control algorithm. Section 4 presents experimental results; discussion follows.

2 Related Work

A water supply network is a complex system, composed of many interconnected elements, which delivers water from reservoirs to consumers. Raw water is generally gathered from a natural source such as a lake, a river or an underground aquifer, and it is then moved to water treatment facilities, where it is purified, and finally the *Water Distribution Network* (*WDN*) delivers water to the demand points. The delivered water must meet the quantity and the quality required by users, in particular the water must reach each part of the network in the specified quantity and under a satisfactory pressure. In order to meet these requirements the water needs to be pushed by hydraulic pumps and it is usually stored in tanks present in the network. The hydraulic pumps require energy to work and water distribution represents a significant percentage of the total cost of the whole water management cycle. For this reason, minimizing pumping costs, while guaranteeing the quality of the service, is one of most widely explored problems in this field. WDN management issues are increasingly important for water utilities managers Bello et al. [2019]: from an operational perspective, the first step to enable network control consists in acquiring field measurements. Sensors and actuators are then installed in the system, to actively intervene in the distribution process, according to the collected measurements and to the model interpreting them. Network control can be automated at two different levels. It can be, in other words, supervised or fully automatic. In the first case, automatic controllers propose corrections to supervising operators, who can still modify them based on their expertise, while in the second case, control actions are applied directly, with no human intermediation. Due to the fact that WDN systems are highly critical, the second solution is probably harder to pursue, at least initially.

Optimizing a water distribution networks is a complex problem, which can be tackled at different levels. *WDN design optimization* consists in choosing the best solution for planning a new distribution system or for expanding an existing one. An exhaustive review of the methods employed for this type of problem is provided by Mala-Jetmarova et al. [2018]. *WDN operation optimization* works

at runtime level, i.e. it is aimed at selecting the best solution for pump control in an existing network. If the control of the system is performed with short time intervals (e.g., some minutes) occurring between two successive correction actions, the WDN is said to be controlled dynamically in real time, as described in Campisano et al. [2010]. In both cases, the choice must minimize the energy, or the cost of energy consumed by pumps, while ensuring that the network is able to meet consumer requirements. The WDN operation optimization can produce a fixed scheduling of pump operations over a time period (e.g. a day or a week) or in continuously updated control settings for the current time step (and possibly the for next future steps). If scheduling is fixed, the problem is called *Pump Scheduling Optimization*.

On the contrary, if the optimal solution is continuously updated based on the most recent data the problem is called *Real Time Pump Optimization*. In this case the optimization problem should be solved at each time step using data gathered by *Supervisory Control and Data Acquisition* (*SCADA*) facilities. Since it is a real-time application, an efficient retrieval of a (near) optimal solution is mandatory in this case. Given the observations gathered on the field, optimization techniques are used to find the best solution compliant with constraints. Since the effort for retrieving the best solution could be unsustainable, meaheuristic approaches can be used, aimed at exploring only a subspace of possible solutions. Metaheuristic approaches are usually combined with a hydraulic simulator which, at each iteration of the algorithm, tests the quality of the explored solution. For real-time optimization when the use of hydraulic simulators is computationally expensive, machine learning methods can be used to in place of the hydraulic simulator, as shown in Rao and Alvarruiz [2007].

In real-world applications, the most used metaheuristic search methods are *Genetic Algorithms* (*GA*), which are inspired by Darwin's theory on the biological processes of reproduction and natural selection. Since a GA simulates a biological process, it includes some randomness and different runs could lead to non-identical results: consequently, there is no guarantee that GA reaches the optimal solution and it may stop at a local minimum. Nevertheless, some experiments have shown that, if the algorithm parameters are chosen correctly, the use of GAs could lead to near-optimal solutions, as shown in Rylander and Gotshall [2002]. Basically, all the variants of GAs are based on the iterative generation of possible solution vectors and on their evaluation based on a specific fitness score.

3 Rule Based Control

3.1 Modeling

As mentioned in Sect. 2, optimization methods require the management of several closed-form differential equations encompassing the physical laws governing the network or, in alternative, the use of hydraulic simulators that implicitly handle these relations. The method proposed in this work, called *Rule Based Control* (*RBC*), requires neither the physical model nor the simulator since the

relationships among parameters can be implicitly extracted from historical data. RBC just needs a set of examples regarding the configurations used in the past and their effectiveness in ensuring the quality of the WDS; given this, RBC is able to act on controllable variables to improve a configuration quickly and limiting the unnecessary control actions. Unlike the other methods, RBC makes it possible to progressively improve a given solution using only a limited set of examples on the problem at hand. It potentially does not even require the definition of a cost function, since a set of solutions labeled as "good" or "bad" is sufficient. These factors make RBC suitable for the optimization of a dynamic system, in particular when

- the evolution of the network cannot be predicted either through closed differential equations or through an accurate hydraulic simulator, or
- the optimization process must be performed real-time and the time required for solving the differential equations or running the simulator is excessive.

The proposed method is, to the best of our knowledge, the only method that requires neither an optimization model, nor a hydraulic model of the WDN that must be optimized.

The definition of a "good solution" depends of course on the requirements and on the boundary conditions of the network; it is, however, possible to define the current performance of the network based on historical statistics. Suppose that a set $\mathcal{S}_H = \{(\boldsymbol{x}_i, y_i)\}$ of historical data is available, where y_i is the observed cost (e.g. the energy consumption per quantity of pumped water) for the configuration characterized by parameter vector $\boldsymbol{x}_i$. Then we can divide historical (as well as present and future) data into two subsets: the "good" configurations and the "bad" ones. In particular, solution $\boldsymbol{x}_i$ is "good" with respect to historical data S_H if $y_i \leq Q_y(S_H, q)$, where $Q_y(S_H, q)$ is the percentile q (where $0 \leq q \leq 1$) of observed cost y_i in historical data S_H (the percentile q is the minimum observed value in data that is greater or equal to a fraction q of all the observed values). For example, by setting $q = 0.5$, a solution is considered good if its indicator is less or equal to the median of the cost of past solutions. Since the problems include some operational constraints, for the WDN optimization the following indicators must be taken into account:

- the *efficiency indicator* y^{eff} that represents the observed value of the objective that must be optimized;
- the *quality indicator* y^{qlt} that represents the observed value of a penalty function.

Since the quality indicator is primary for a good solution, it is possible to consider a solution $\boldsymbol{x}_i$ good with respect to historical data $\mathcal{S}_H$, only if both the following relations hold:

$$y_i^{\text{qlt}} \leq Q_{y^{\text{qlt}}}(\mathcal{S}_H, q) \tag{1}$$

$$y_i^{\text{eff}} \leq Q_{y^{\text{eff}}}(\mathcal{S}_H^*, q) \tag{2}$$

where $\mathcal{S}_H^*$ is the subset of the $\mathcal{S}_H$ satisfying condition 1. In other words, to be considered good, a solution must have both a good quality indicator and a good efficiency indicator. This leads to consider a single discrete indicator, with 3 possible classes: (a) *bad quality of the service* (BadQoS), the worst class indicating that the constraints are poorly respected; (b) *bad energy efficiency* (BadEff), the intermediate class and indicating that the constraints are respected but the cost function is too high; (c) *good energy efficiency* (GoodEff), the best class in which the constraints are respected and the efficiency indicator is good. In particular, it is possible to assign the discrete output value y_i^{st} to a configuration $\boldsymbol{x}_i$ as follows:

$$y_i^{\text{st}} = \begin{cases} \text{BadQoS} & \text{if } y_i^{\text{qlt}} > Q_{y^{\text{qlt}}}(\mathcal{S}_H, q) \\ \text{BadEff} & \text{if } y_i^{\text{qlt}} \leq Q_{y^{\text{qlt}}}(\mathcal{S}_H, q) \text{ and } y_i^{\text{eff}} > Q_{y^{\text{eff}}}(\mathcal{S}_H^*, q) \\ \text{GoodEff} & \text{if } y_i^{\text{qlt}} \leq Q_{y^{\text{qlt}}}(\mathcal{S}_H, q) \text{ and } y_i^{\text{eff}} \leq Q_{y^{\text{eff}}}(\mathcal{S}_H^*, q) \end{cases} \tag{3}$$

Notice that Eq. 3 represents only one of the possible definitions of y_i^{st} as the classes can be defined arbitrarily depending on the problem at hand. In particular, operational constraints on the pumps or any other prior knowledge about the WDN can be inserted as well. This first step is called *output definition* and it produces a dataset $S = \{(\boldsymbol{x}_i, y_i^{\text{st}})\}_{i=1}^N$ with N examples, that can be used as a training set for the learning phase of a rule-based classification model. The learning phase represents the second step of the optimization process and is called *rule generation*. It consists in using historical data to generate the model that describes the relationship between inputs (parameters of the WDN) and output (performance of the WDN). For the rule generation step the use of *Logic Learning Machine* (*LLM*) method is suggested Muselli [2006]. Notice that the rule generation phase could be done periodically (e.g. every day) in order to take into account changes in the behavior of the system. Moreover, the rules are built allowing overlaps in the input space so that in some values is missing (e.g. because a sensor is not sending information), other rules, involving other sensors can be used, instead.

To allow an effective evaluation of the set of rules $\mathcal{R}$ it is possible to associate some quality measures with each of them. A generic rule $r \in \mathcal{R}$ can be seen as a couple $(\boldsymbol{\kappa}(r), O(r))$ where $\boldsymbol{\kappa}(r)$ is a vector of conditions, while $O(r)$ is the output in its consequence. The attribute associated with a condition κ is denoted with $A(\kappa)$. An example $\boldsymbol{x}$ is said to be *covered* by r (denoted by $\boldsymbol{x} \leq r$) if it satisfies all the conditions in $\boldsymbol{\kappa}(r)$. Vice versa, $\boldsymbol{x} \not\leq r$ means that $\boldsymbol{x}$ is not covered by $\kappa(r)$. The *covering* $\chi(r)$ of r is the fraction of cases $(\boldsymbol{x}_i, y_i)$ in the training set having $y_i = O(r)$ and input $\boldsymbol{x}_i$ covered by r while the *error* $\epsilon(r)$ of r is the fraction of example in the training set having output $y_i \neq O(r)$ covered by r.

Combining covering and error we can define the *relevance* of a rule as

$$R(r) = \frac{N_{O(r),\boldsymbol{\kappa}(r)}}{N_{O(r)}} \left(1 - \frac{N_{\overline{O(r)},\boldsymbol{\kappa}(r)}}{N_{\overline{O(r)}}}\right)$$

The relevance of rules can be used to determine the output associated with a generic input vector $\boldsymbol{x}$. To this aim, a score $S(\boldsymbol{x}, c) = \sum_{r \in \mathcal{R}_{\boldsymbol{x}}^c} R(r)$ can be

computed for each output class c, where $\mathcal{R}_{\boldsymbol{x}}^{c}$ is the subset of rules covering $\boldsymbol{x}$ with an output of $O(r) = c$. The input vector $\boldsymbol{x}$ will then be associated with the class c scoring the highest value of $S(\boldsymbol{x}, c)$.

3.2 Control

After the rule-based model is generated, a third step called Rule Based Control is continuously applied: this task takes the current configuration $\boldsymbol{x}$ as input and it generates a new solution by making changes in order to improve $\boldsymbol{x}$. RBC finds all the rules that the current configuration $\boldsymbol{x}$ could match and then changes some values in $\boldsymbol{x}$ to satisfy other good rules. In general, it is worth noting that $\boldsymbol{x}$ can contain not only the controllable decision variables but also parameters that are non-controllable (e.g. time). RBC is obviously configured to change only those values related to controllable variables.

The RBC procedure is applied at each time step until a termination condition is met and the produced solution becomes part of the historical data. In particular step 1 and 2 (output definition and rule extraction) must be performed only when the changes applied in the Rule Based Control step no longer lead to improvements or the classification model performance drops on average over a specific time period.

Before applying RBC, a weight value should be associated with each output class so that the higher the weight the higher the cost of the class. In particular negative weights should be associated with target classes, while positive weights should be assigned to unwanted outputs. Moreover, weights should be chosen so that $w_{\text{BadQoS}} \geq w_{\text{BadEff}} \geq w_{\text{GoodEff}}$. The goal of the optimization procedure is to find the vector $\boldsymbol{x}$ that minimizes the probabilities $P(\text{BadQoS} \mid \boldsymbol{x})$ and $P(\text{BadEff} \mid \boldsymbol{x})$ and maximizes the probability $P(\text{GoodEff} \mid \boldsymbol{x})$ for each WDN configuration $\boldsymbol{x}$. Since the probabilities $P(c \mid \boldsymbol{x})$ are usually unknown a priori, they can be estimated using the rules derived from LLM. The generic probabilities $P(c \mid \boldsymbol{x})$ derived from a LLM model can be obtained as a function of the rules covering $\boldsymbol{x}$ relative to c and of the rules covering $\boldsymbol{x}$ relative to classes other than c.

$$P(c \mid \boldsymbol{x}) = \frac{S(\boldsymbol{x}, c)}{\sum_{k \in C} S(\boldsymbol{x}, k)} \tag{4}$$

where $C = \{\text{BadQoS}, \text{BadEff}, \text{GoodEff}\}$ is the set of all possible output classes.

Then, the cost function $g(\boldsymbol{x})$ that RBC must try to minimize is introduced:

$$g(\boldsymbol{x}) = \sum_{c \in C} w_c P(c \mid \boldsymbol{x}) = \frac{\sum_{c \in C} w_c \cdot S(\boldsymbol{x}, c)}{\sum_{k \in C} S(\boldsymbol{x}, k)} \tag{5}$$

From Eq. 5, it is clear that classes with $w_c < 0$ produce a negative contribution to the cost function and therefore the relative score should be increased while the score associated with classes where $w_c > 0$ must be lowered to reduce the cost function g. A class with $w_c = 0$ does not contribute to the cost, so the associated score is not relevant. Each $P(c \mid \boldsymbol{x}_i)$ can be increased (resp. reduced)

by making changes to $\boldsymbol{x}$ in order to add (resp. remove) rules to $\mathcal{R}^c_{\boldsymbol{x}}$. RBC aims to change $\boldsymbol{x}$ so that more rules for the good class and fewer rules for the bad classes are matched. Given the set E of modifiable attributes (e.g. the number of active pumps), RBC (see Algorithm 1) is building the set $\mathcal{R}_{\boldsymbol{x},E}$ of rules that could potentially cover $\boldsymbol{x}$ after some changes in attributes belonging to E. In other words, the $\mathcal{R}_{\boldsymbol{x},E}$ set contains the rules that matches $\boldsymbol{x}$ at least regarding the non-modifiable attributes, such as the time of the day. Similarly, the sets $\mathcal{R}^+_{\boldsymbol{x},E}$ and $\mathcal{R}^-_{\boldsymbol{x},E}$ of good and bad rules respectively, which could potentially cover $\boldsymbol{x}$, can be defined as well. The algorithm is basically searching inside $\mathcal{R}^+_{\boldsymbol{x},E}$ considering the rules starting from the class with the lowest weight (i.e. the most desirable class, if there is more than one class with $w_c < 0$). The selection is actually performed by finding the maximum of $\big(-w_{O(r)}, R(r)\big)$ according to lexicographic ordering: first the weight of the output class and then, in case of a tie, the relevance of the rule, is considered. Starting from the selected rule, the algorithm makes the minimum changes in $\boldsymbol{x}$ (creating a modified vector $\boldsymbol{x}^M$) so that r covers $\boldsymbol{x}^M$; once it is covered by r, the value of $g(\boldsymbol{x}^M)$ is computed and if it is smaller than $g(\boldsymbol{x})$, the changes are maintained, otherwise they are dropped.

Data: $\boldsymbol{x}$, $\mathcal{R}_{\boldsymbol{x},E}$, $\boldsymbol{w}$, E
Result: $\boldsymbol{x}^M$
$\mathcal{R}^+_{\boldsymbol{x},E} = \{r \in \mathcal{R}_{\boldsymbol{x},E}$ such that $w_{O(r)} < 0\}$
while $\mathcal{R}^+_{\boldsymbol{x},E}$ is not empty **do**
  $r^* = \arg\max_{r \in \mathcal{R}_{\boldsymbol{x},E}} \big(-w_{O(r)}, R(r)\big)$
  $\hat{\boldsymbol{\kappa}}(r^*) = \{\kappa \in \boldsymbol{\kappa}(r^*)$ such that $A(\kappa) \in E\}$
  $\boldsymbol{x}^M = \boldsymbol{x}$
  for $\kappa \in \hat{\boldsymbol{\kappa}}(r^*)$ **do**
    $j = A(\kappa)$
    if $x_j \not\leq \kappa$ **then**
      $x^M_j = \arg\min_v \{\| x_j - v \|$ such that $v \leq \kappa\}$
    end
  end
  if $g(\boldsymbol{x}^M) < g(\boldsymbol{x})$ **then**
    $\boldsymbol{x} = \boldsymbol{x}^M$
  end
  $\mathcal{R}^+_{\boldsymbol{x},E} = \mathcal{R}^+_{\boldsymbol{x},E} \setminus \{r^*\}$
end

Algorithm 1: Function **RuleBasedControl** that implements the control strategy based on rules

It is worth noting that since it is possible that the optimum cannot be reached, Algorithm 1 is not the best approach. In this situation, avoiding the worst configuration (i.e. BadQoS) could bring to better results. This can be done by applying RBC in a hierarchical way, leading to the *Hierarchical Rule Based Control* (*HRBC*) algorithm. In particular, in the first application only the top class (i.e. GoodEff) has a negative weight; at the subsequent step if

the best output cannot be reached, then a new class (i.e. BadEff) is set with a negative weight. In general, the goal of the first application of RBC is to move the considered solution towards the best class. If the goal is not achieved, the goal becomes less ambitious and consequently consists in moving the considered solution towards an acceptable class; at every step the set of acceptable classes is enriched and the goal is less restrictive. Each step of the algorithm consists in applying RBC, then only if the predicted output does not reach the target classes at the last step, the weights are updated and the step is repeated. The score update consists in setting the minimum positive weight equal to the maximum between the current negative weights. Also, all the negative weights are decreased.

4 Results

In order to verify the effectiveness of the RBC method, different tests have been performed using simulated data of a fictitious water distribution system named D-Town, whose hydraulic model was developed by the University of Exeter and is available online Salomons et al. [2012]. The D-Town hydraulic model consists of 399 junctions, 7 storage tanks, 443 pipes and 11 pumps divided into 5 pumping stations. This network has also been the object of *The Battle of the Water Networks II*, discussed by Monteiro et al. [2014], a competition that required the optimization of the network design to cope with the increased water demand of the population.

The dataset to train and to evaluate the model was generated using *EPANETTOOLS*, a package enabling the user to call all the EPANET programmer toolkit functions by means of Python scripts. After collecting data and defining the output, a model selection phase was performed in order to choose the optimal LLM parameters for generating the rule-based model. A training set composed by $67K$ samples was generated, by simulating an initial state for pumps and tanks, as well as a water demand profile and evaluating the quality of service and efficiency on top of that. Referring to 3.1, the QoS is, more specifically, computed by observing whether, and by how much, the measured pressure is lower than the lower limit of the optimal pressure range or greater than the higher extreme of the optimal pressure range. The Eff is computed by measuring the energy consumption of the pumps.

The LLM is then used to extract rules using the level of the tanks and the status of the pumps to predict an output value. As described in 3.1 three possible output values are defined: BadQoS, BadEff and GoodEff. The model trained by LLM obtains an accuracy of 0.84 and a Cohen's K of 0.76. In order to measure the effectiveness of the control actions suggested by the RBC approach, two terms of comparison are considered. Firstly, the behavior of the system after applying the controls is compared with the behavior of the system if no actions had been taken. We will refer to this as the *No Action* case. A competing approach is also used to suggest control actions, by applying *Genetic Algorithms* to the same problem. In all cases, the output of the system in the next time step

(i.e., after 15 min) is evaluated by the EPANET simulator. The RBC approach does not need a simulator and can rely on the Logic Learning Machine model to determine the control actions to be suggested. Genetic Algorithms, on the other hand, require the use of a simulator within their optimization pipeline, to find the best control actions to be suggested. This implies that the adoption of Genetic Algorithms is subordinated to the formulation of a physical model of the network. This model will estimate, with some degree of approximation, the behavior of the system. To set up the comparison, EPANET is used as a support for Genetic Algorithms, but Gaussian noise is added to the supplied water demand profile, in order to also take into account the approximation which characterizes the physical model of the system. In other words, the real water main has a demand ranging from 80% to 120% of the expected demand obtained through the simulator and used for the optimization. In the tests the three scenarios (no action, controls by RBC, controls by GA) are taken into consideration and compared using the $30K$ samples in the test set, checking their evolution after 15 min when the different control strategies are applied.

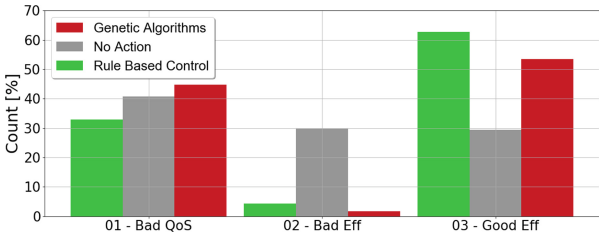


Fig. 1. The performance of the three approaches considering the indicator that takes into account both quality of the service and energy optimization.

Considering both QoS and energy efficiency, it is possible to see (in Fig. 1) that, referring to the BadQoS, BadEff and GoodEff possible output values, RBC gets the lowest percentage of cases in the worst class (BadQoS) and the highest percentage of cases in the best class (GoodEff). It is also worth considering that turning pumps on and off at a high rate in such a complex and critical system as a water main may not be desirable, for stability reasons. To take this aspect into account, RBC and GA have been compared as regards the number of proposed switches, i.e. changes on/off of the pumps. In summary, the average number of changes per iteration for RBC is 2.25, compared to 4.24 for GAs. In particular, in more than 60% of cases RBC suggests 0, 1 or 2 switches, while GAs propose 2, 4 or 5 changes in more than 60% of the situations. This improvements lead to a 5% of estimated energy saving, while keeping the same level of quality of service. As concerns computation time, on average, RBC requires 0.04 s for a single run, while a single run for the GAs requires 56 s. Computational efficiency, combined with the fact that RBC does not require a mathematical model of the network, makes it particularly well-suited for a reactive, on-line control application.

5 Discussion

Rule Based Control, a new rule-based machine learning method for optimizing complex networks, has been presented in this work. In particular, we focused on applying this method to pump operation optimization in water distribution networks, where it is important to save energy and maintain a good quality of the service. However, the same approach can be applied in a similar way to optimizing the control of other systems. A limitation of the Rule Based Control method consists in the fact that training data must be available, so that the underlying classification algorithm can learn about the behavior of the target system. The test consisted in using a hydraulic simulator to measure how the new approach can improve the quality of the network. The test performed on a separate test set highlighted that Rule Based Control is able to improve the network both in terms of quality of service and energy efficiency. Moreover the comparison with a more traditional approach based on Genetic Algorithms pointed out that the new approach is able to (a) reach higher levels of network improvement (b) reduce the number of switches in pumps, which should result in lower maintenance costs for the water main (c) reduce the computation time, thus enabling real time applications.

References

Bello, O., Abu-Mahfouz, A.M., Hamam, Y., Page, P.R., Adedeji, K.B., Piller, O.: Solving management problems in water distribution networks: A survey of approaches and mathematical models. Water **11**(3), 562 (2019)

Campisano, A., Creaco, E., Modica, C.: Rtc of valves for leakage reduction in water supply networks. J. Water Resour. Plan. Manag. **136**(1), 138–141 (2010)

Ghorbanian, V., Karney, B., Guo, Y.: Pressure standards in water distribution systems: reflection on current practice with consideration of some unresolved issues. J. Water Resour. Plan. Manag. **142**(8), 04016023 (2016)

Mala-Jetmarova, H., Sultanova, N., Savic, D.: Lost in optimisation of water distribution systems? a literature review of system design. Water **10**(3), 307 (2018)

Monteiro, A.J., et al.: Battle of the water networks II. J. Water Resour. Plan. Manag. **140** (2014)

Muselli, M.: Switching neural networks: a new connectionist model for classification. In: Apolloni, B., Marinaro, M., Nicosia, G., Tagliaferri, R. (eds.) NAIS/WIRN -2005. LNCS, vol. 3931, pp. 23–30. Springer, Heidelberg (2006). https://doi.org/10.1007/11731177_4

Rao, Z., Alvarruiz, F.: Use of an artificial neural network to capture the domain knowledge of a conventional hydraulic simulation model. J. Hydroinf. **9**, 15–24 (2007)

Rylander, S., Gotshall, B.: Optimal population size and the genetic algorithm. In: World Scientific and Engineering Academy and Society (WSEAS) international conference on Soft Computing, Optimization, Simulation, and Manufacturing Systems (SOSM 2002) (2002)

Salomons, E., Ostfeld, A., Kapelan, Z., Zecchin, A., Marchi, A., Simpson, A.R.: The battle of water networks ii. In: Water Distribution Systems Analysis Conference 2012 (2012)

Graph-Based Segmentation and Markov Random Field for Covid-19 Infection in Lung CT Volumes

Giulia Iaconi(✉), Federica Ferraro, Marco Balletto, David Solarna, Marco Trombini, Gabriele Moser, and Silvana Dellepiane

Electrical, Electronics and Telecommunication Engineering and Naval Architecture Department (DITEN), Università degli Studi di Genova, 16145 Genoa, Italy
{giulia.iaconi,federica.ferraro,david.solarna, marco.trombini}@edu.unige.it, {gabriele.moser, silvana.dellepiane}@unige.it

Abstract. Most of the papers in the literature concerning the detection and evaluation of particular features from medical images (i.e., trauma, infections, etc.) are based on supervised techniques that always require large amounts of annotated data. On the contrary, in this work a completely unsupervised method is proposed, inspired by a previous 2D approach to segmentation, reformulated and extended to the case of 3D tomographic images.

The method uses a flooding algorithm for the extraction of regions in the lung area, followed by the association with the corresponding statistical model by parametric estimation. A maximum a posteriori estimate of the 3D label map is obtained by modeling the image as a Markov Random Field that includes space-contextual information by means of a simple local proximity model for inter-voxel dependence. At this level, the problem is reformulated in terms of energy minimization, which is addressed using the α - β swap graph-cut algorithm.

The proposed method was tested on 3D lung images in order to delineate infections from CT scans of the chest. A set of COVID-19 confirmed patients was then considered, and the results obtained showed accurate detection ability and remarkable robustness to the heterogeneity of CT scans considered.

Keywords: Unsupervised segmentation · Graph-cut · Biomedical image · Markov random field

1 Introduction

Diagnosis is the process by which it is possible to identify the type of pathology or disease affecting the patient. To obtain the characteristics associated with different diseases and make it accessible to the medical expert for further evaluation, imaging techniques such as ultrasound, digital radiography, and computerized tomography (CT) can be used.

Many different automated infection detection and assessment techniques have already been proposed and validated in the literature, often based on machine learning techniques and, in particular, on various deep neural network (DNN) architectures.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 43–52, 2023.
https://doi.org/10.1007/978-3-031-16281-7_5

The application of DNN to X-ray and CT data is a well-known approach whose capabilities in terms of object detection, localization, classification and semantic segmentation for medical purposes have been widely discussed, as reported in [1].

With respect to the specific case of chest radiography and tomography, DNNs have been already employed, for example, to verify the presence and identify the nature of nodules and inflamed areas for lung cancer and Interstitial Lung Disease (ILD) diagnosis [2–4]. In other works, however, the focus is shifted on the preliminary phases of image pre-processing and conditioning. In these cases, the aim is to develop solutions helping the process of localization and identification of the elements and details of the image that have relevance for diagnostic purposes. Although the generation of the diagnosis is still delegated to the expert, such tools are employed to provide additional support, making the whole process faster and more consistent.

One of the most common ways to support the work of the expert is by providing a segmentation tool. In *Qi et al.* [5], for example, a U-Net model is employed for the segmentation of the main regions of the lung and for the segmentation of the infection patches in each region. For this purpose, various supervised solutions have been proposed in the literature, suitably designed to learn how to perform segmentation automatically based on the type of application, for example to detect lung regions associated with a certain type of lesion. In addition, the quantitative analysis of the extracted areas and volumes allows the computation of various severity indexes that provide to the operator additional indications for the assessment of the disease.

Up to now, in the clinical practice, this type of analysis has been mainly carried out manually by the expert. However, it is tedious, time consuming, and may become problematic in all those situations where the number of patients increases rapidly. As a consequence, great importance is given to the development of tools for an automatic analysis of the imagery that can provide support and reduce the level of effort required to the expert.

In the present paper, a method is proposed to perform an automatic segmentation of the main regions of interest from lung tomography as an aid to the expert in the detection and localization of diagnostic findings relevant to the COVID-19 diagnosis.

The novel method proposed here is mainly based on the studies performed by *Gemme et al.* [6] for graph-based segmentation, and *Moser et al.* [7] for Markov Random Field (MRF) regularization, specifically revised and adapted to 3D CT scans.

A preliminary phase of approximate segmentation is performed on the basis of topological information, followed by a phase devoted to estimate the parametric region models, and a subsequent MRF contextual analysis. To this end, the proposed graph-based segmentation, following the description given in [6], makes use of a flood-filling approach instead of a Spectral Graph cut to take into account at a time topological and opacity information. The final MRF step allows the exploitation of the statistical models in spatially contextual energetic terms.

As an advantage, the current unsupervised approach does not require any training sets and, as a result, the expert does not have to go through an onerous and time-consuming process of manually annotating a large 3D database. Furthermore, in this way it is possible to extract, without supervision, elements of interest from 3-D images to help the annotation required in the creation of large databases.

2 Materials and Methods

2.1 Overview

In this work, the segmentation activity is carried out through a sequence of processing steps. At the beginning, in the pre-processing phase, a mask is automatically generated separating the lung voxels from the background to reduce the size of the elements involved in each phase and make the whole procedure computationally lighter.

The second processing phase involves a Region Growing algorithm starting from a set of randomly positioned seed points for the extraction of different lung regions from the CT scan, the so-called Regions of Interest (ROI). Then, a conditional statistical model of the region is constructed for each ROI by fitting a Gaussian mixture to the extracted samples.

The segmentation problem is then expressed in a Bayesian framework in terms of a probabilistic Markov Random Field model since a contextual/spatial regularization approach is essential to transform the global maximization problem into a more tractable one of energy minimization. The approximate solution is obtained in terms of cutting the graph using the $\alpha - \beta$ *swap* algorithm.

2.2 Dataset

The dataset is composed by a total of 20 CT scans of confirmed COVID-19 patients and derives from the combination of the data contained in two different online repositories [8, 9], namely *"Coronacases"* and *"Radiopaedia"*. All scans in the dataset are equipped with lung and infection masks annotated by an experienced radiologist, useful for describing the nature of the findings, which will be used as a ground-truth reference during evaluation phase.

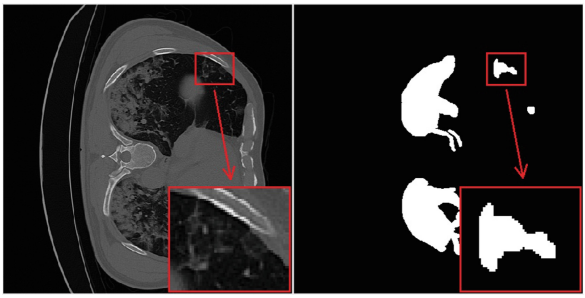


Fig. 1. An example section of a CT scan (left) and the corresponding infection mask image (right) from online repositories.

Starting from the spatial sequence, the masks are constructed through a process of manual segmentation of the areas of interest (i.e., the lung parenchyma and the regions showing non-negligible opacification). As shown in Fig. 1, this process is inevitably affected by approximations since the delineation of the mask tends to overfill some areas that do not appear uniformly opaque in the given scan.

2.3 Lung Masking and Region Segmentation

A raw CT scan is taken as input, where radio-density is measured, as usual, in Hounsfield Units (HU). After a thresholding step, a Mathematical Morphology Closing operator generates the binary mask, which separates the lung volume from the background. N is the size of the lung volume.

We can then proceed with the segmentation applied to the lung volume, with the aim of delineating the most significant Regions of Interest starting from a random set of seed points, S. Referring to the method described in [6], the graph-based ROI segmentation is based on the computation of the cost functions $F = \{F_n\}$, each associated with a specific seed $s_n \in S$.

Let the tomographic volume X be mapped into a non-empty, fully connected, undirected and vertex-weighted graph $G = (I, E)$, where I is the set of vertices representing the voxels, and E is the set of arcs linking voxels pairs. For each couple of nodes i and j, the difference function, w, is defined as:

$$w : I \times I \rightarrow R, w(i, j) = |x_i - x_j| \tag{1}$$

In Eq. 1, the higher the value of $w(i, j)$, the more different the two nodes are in terms of gray level.

By fixing a specific node s_n belonging to the graph, such a distance is then computed with respect to each voxel, $w_n(i) = w(i, s_n), \forall i \in I$, and thus each node in the graph, is associated with a value representing its dissimilarity with respect to s_n. The computation of $F_n(i), \forall i \in I$, given the seed $s_n \in S$, is obtained through the flood-fill algorithm [10].

2.4 Parametric Model Fitting

The main consequence for the ROI extraction is the possibility of estimating the probabilistic models that could allow the calculation of the likelihoods for the generation of a possible final semantic classification/segmentation labeling. In particular, the voxels that make up a ROI are used to obtain an estimate of the probability density function (pdf) of the variable x measured in Hounsfield Units (HU), conditional on that region. The parametric statistical model, $g(\theta, x)$, is based on the Gaussian mixture. At the end of the model adaptation phase, a pdf estimate is available for each of the extracted ROIs.

2.5 Markov Random Field Modelling

Once the parametric statistical model has been defined, it is possible to calculate the probability that all the voxels in the lung region are part of one of the extracted ROIs.

Indicating with $X^* \in R^N$ the vector of all the N observations in the lung region and with $Y \in Y^N$ the random vector of the N labels composing the output map, the Maximum A posteriori Probability (MAP) criterion could be applied to find the best configuration of labels Y^* that maximizes the posterior probability mass function $p(Y|X = X^*)$:

$$Y^* = \text{argmax}_{Y \in Y^N} p(Y|X = X^*) \tag{2}$$

In this way, the choice of the label to be assigned to each voxel is performed in a spatial contextual framework, by considering the whole lung volume.

As is well known, this approach is highly impractical due to the total number of different configurations that should be evaluated (NP-complete problem). By indicating with M the cardinality of the set Y, and being N the number of voxels to be labeled, the quantity N^M of the possible label maps becomes extremely large even for small values of M. This is why a 3D Markov Random Field (MRF) [10] model has been adopted to describe the inter-voxel dependence so that the global optimization problem becomes tractable.

A first order 3D neighborhood system is adopted generating very simple cliques as shown in Fig. 2.

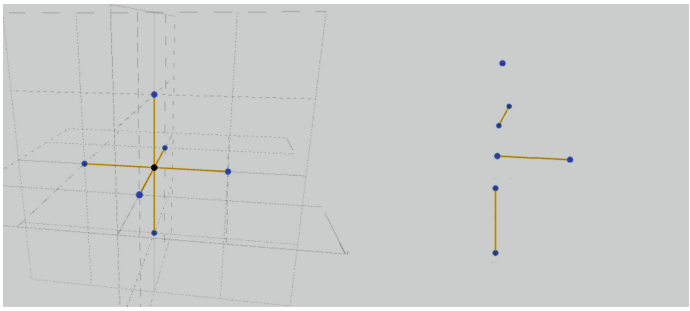


Fig. 2. First order neighborhood system (on the left) and the four corresponding cliques (on the right).

Using the notation $j \sim i$ to indicate that j is a neighbour of the element i with respect to a certain 3D neighborhood system, the Markov condition can be written as:

$$p\left(y_i|y_j, \forall j \neq i\right) = p\left(y_i|y_j, j \sim i\right). \tag{3}$$

Such condition implies that, for the choice of the optimal label y_i*, the totality of the elements in the grid carries as much information as the subset of elements in the neighborhood of i. In fact, the statistics of the random variable y_i are conditioned by other labels only if they belong to the neighborhood. Modelling the spatial-contextual correlation by means of a MRF allows to pass from the global joint maximization of the posterior mass function $p(Y|X)$ to a neighbour-local formulation of the problem.

The MRF hypothesis has a crucial impact on the problem of mapping. Indeed, thanks to the so-called Hammersley–Clifford theorem [11, 12], it is possible to prove that, under mild assumptions, if Y is an MRF, then the global posterior probability can be written as the following Gibbs distribution:

$$P(Y|X) = \frac{e^{-U(Y|X)}}{Z(X)} \tag{4}$$

where the function $U(Y|X)$, called energy function, is defined locally according to the neighborhood system. $Z(X)$, usually named "partition function", is a normalization factor [13]. Therefore, the maximization of the global posterior probability mass function $P(Y|X)$ is equivalent to the minimization of the energy function $U(Y|X)$. Thanks to

the local neighborhood-based definition of the energy, this minimization problem is computationally tractable [12, 13], a remarkable result that offers a feasible solution and explains the widespread success of MRF-based approach.

In this case, the technique chosen to make the solution computationally feasible is the α - β *swap* algorithm [15], which reformulates the problem as a sequence of binary subproblems and, for each subproblem, a global energy minimum is obtained through the max-flow/min-cut formulation.

3 Results

The actual evaluation of the performances achieved by the proposed method is expressed in terms of the ability to detect the boundaries that delineate the separation between lung parenchyma and disease areas.

The use of the first order neighborhood system, together with an appropriate pair potential (favoring the continuity between adjacent labels), allows to characterize the spatial dependence in an appropriate way for the type of images being analyzed.

In fact, the aim of the solution presented in this work is to provide support to the expert by localizing and segmenting suspect ROIs in a lung CT volume. In this respect, of fundamental importance is the set of infection masks obtained for each raw scan (Fig. 3). Such masks, manually built by an expert radiologist, provide information about any finding of actual diagnostic interest in the images and can be employed as reference ground truth for testing.

A first qualitative analysis, performed by means of a simple visual comparison, shows good correspondence between the areas labelled by the expert and the ROIs appearing in the output label map in Fig. 3. As for the quantitative evaluation of the results, the overall sensitivity is used to check what percentage of the ground-truth infection voxels is covered by the output ROIs (i.e., true positives).

The higher the sensitivity, the lower the percentage of infection that is confused with the background of the healthy lung and, consequently, remains undetectable for the diagnosis phase (i.e., missed alarm errors).

Fig. 3. Example of the correspondence that can be obtained between true-infection (in white in the left image portion) and output ROIs (in gray in the right image portion).

The values reported in Table 1 show an average detection sensitivity of around 75% computed for each scan of online repositories. In the 7th database Radiopaedia case, no value is reported as the number of ground-truth infection voxels is so small that the sensitivity is not meaningful. However, it should be noted that the actual sensitivity could be higher than that indicated in the table. In fact, in many cases, the infection masks, built manually by the expert with imprecise image editing tools, extend too far around the area of the infection and even include some uninteresting voxels.

Table 1. Table of the overall sensitivities achieved for the infected areas for each scan of the dataset.

	Repository	
Patient	Coronacases	Radiopaedia
1	0.51	0.74
2	0.88	0.80
3	0.91	0.86
4	0.68	0.90
5	0.93	0.62
6	0.59	0.80
7	0.92	N.A
8	0.80	0.65
9	0.69	0.50
10	0.50	0.91

When merged (correctly) with the lung background, these voxels will be (erroneously) considered as missed alarms, thus contributing to the reduction in measured sensitivity. For some CT scans, very good results are obtained by considering label-by-label sensitivity, namely, by measuring how many voxels of the total infection are assigned to label y_i. In this way, it is possible to check whether there are labels in Y that cover most of the infected areas. Once the most important labels have been identified, precision can be considered by measuring how many of the y_i labels in the output map are assigned to infected voxels. This index allows the accuracy with which the y_i label is able to identify infection to be verified.

In this case (Fig. 4), only 10% of the total infected volume blends into the background (black), and the number of missed alerts is almost negligible. More than 73% of labeled pixels is covered by only two classes, which are depicted in green and orange. Moreover, very high precision is obtained for both classes (about 95% and 90%, respectively), which means that very few voxels labeled green or orange identify anything other than an infection.

Fig. 4. Raw CT scan with the associated ground-truth infection mask and the corresponding output label map (right) that shows the two classes of infected voxels (orange and green).

It is worth noting, in this case, how the proposed method is able to detect the opacity even when it is very slight and its boundary with the background region is not very neat. However, it may be the case that the value for overall sensitivity may be strangely low, despite the fact that there is a visual correlation between unconsolidated opacities and segmented ROIs. This is not necessarily an indication of poor performances for the proposed method, but it may depend on the imperfections of infection mask. For instance, Fig. 5, shows a case in which the proposed method obtains a meaningful discrimination between opacity and background.

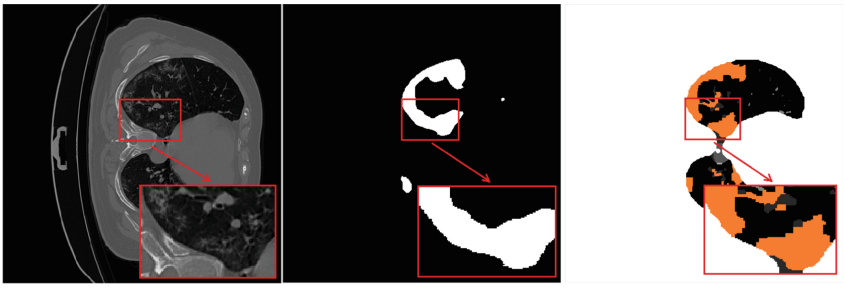


Fig. 5. Example of the higher precision achievable by the proposed method (right) as compared to the manually labelled mask (center). The details in the red boxes show how, in the latter case, many non-infection voxels are marked as well.

4 Conclusion

In the context of the COVID-19 pandemic, rapid and accurate diagnostic tools are essential both to track the spread of the disease and to assess the severity of the infection in order to identify adequate therapy.

In this work a solution has been proposed with the ability to provide support to the expert in analyzing the content of CT lung volumes. The presented approach combines methodological ideas drawn from different areas of image processing and modeling,

including 3D graph-based segmentation, 3D Markov Random Fields, and parametric density estimation. One of the most critical aspects that has been investigated in this work is the computational practicability of the proposed method.

In fact, performing the segmentation on the whole volume could be quite heavy due to the size of the involved chest CT scans. To deal with such data, an automatic process of lung masking has been included in this work. Knowing that only a small part of the total imaged volume has a diagnostic relevance for COVID-19, each voxel that falls outside the lung region is discarded. By doing so, it is possible to minimize the amount of data employed at each stage of the process. For each scan, an infection mask is available composed by an expert through a manual labelling of the findings considered diagnostically relevant. Such masks are employed as reference ground-truth data for evaluating the detection performances of the proposed method.

By means of the computation of both sensitivity and precision indexes, but also by qualitative analysis, it was demonstrated that the solution developed in this work is capable of localizing, in average, the 75% of the total infected volume. Moreover, in many cases, the method allows to minimize the percentage of the total detected volume that is not infection-related, thus avoiding confusion and preventing the overestimation of the severity of the disease.

Besides, the potential of the proposed approach, acting in an unsupervised manner, would facilitate the manual work of experts, reduce the variability of evaluators, and close the experience gap among experts (semantic information is provided manually by medical experts and is thus affected by intra-rater and inter-rater variability and the experience of annotators).

Among possible future extensions of this work, it is worth mentioning the development of more sophisticated parametric models for the radio-density distributions for CT scans, able for example to characterize the underlying texture [16], the use of more complex spatial-contextual models involving higher order cliques [17] and the embedding of texture-related features in the Markovian energy function for a better discrimination of the findings with similar radiographic responses.

Finally, the proposed adaptive method that integrates various forms of computational intelligence (e.g., reasoning on graphs, aggregation of regions, statistical models, and contextual analysis) is able to demonstrate that even the unsupervised approach can create good results that could support the phase of medical diagnosis and the creation of large annotated databases.

References

1. Bhattacharya, S., et al.: Deep learning and medical image processing for coronavirus (COVID-19) pandemic: a survey. Sustain. Cities Soc. **65** (2021)
2. Liao, F., Liang, M., Li, Z., Hu, X., Song, S.: Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-or network. IEEE Trans. Neural Netw. Learn. Syst. **30** (2019)
3. Lo, S.-C., Lou, S.-L., Lin, J.-S., Freedman, M., Chien, M., Mun, S.: Artificial convolution neural network techniques and applications for lung nodule detection. IEEE Trans. Med. Imaging **14** (1995)
4. Shin, H.-C., Roth, H., Gao, M., Lu, L., Xu, Z., Nogues, I.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35** (2016)

5. Qi, X., et al.: Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with sars-cov-2 infection: a multicenter study. medRxiv (2020)
6. Gemme, L., Dellepiane, S.G.: An automatic data-driven method for sar image segmentation in sea surface analysis. IEEE Trans. Geosci. Remote Sens. **56**(5), 2633–2646 (2018)
7. Moser, G., Serpico, S.B., Benediktsson, J.A.: Land-cover mapping by markov modeling of spatial–contextual information in very-highresolution remote sensing images. Proc. IEEE **101**(3), 631–651 (2013)
8. Paiva, O.A.: Dataset: Coronacases.org - helping radiologists to help people in more than 100 countries (2020)
9. Glick, Y.: Dataset: Covid-19 pneumonia (2020)
10. Trombini, M., Solarna, D., Moser, G., Dellepiane, S.: A goal-driven unsupervised image segmentation method combining graph-based processing and Markov random fields. Pattern Recogn. (in press)
11. Geman, S., Geman, D.: Stochastic relaxation Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-6 **6**, 721–741 (1984)
12. Dubes, R.C., Jain, A.K.: Random field models in image analysis. J. Appl. Stat. **16**(2), 131–164 (1989)
13. Li, S.Z.: Markov Random Field Modeling in Image Analysis. Springer, New York (2010). https://doi.org/10.1007/978-1-84800-279-1
14. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? IEEE Trans. Pattern Anal. Mach. Intell. **26**(2), 147–159 (2004)
15. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. **23**(11), 1222–1239 (2001)
16. Al-Kadi, O.S.: Combined statistical and model-based texture features for improved image classification. In: 4th IET International Conference on Advances in Medical, Signal and Information Processing (MEDSIP 2008) (2008)
17. Kato, Z., Zerubia, J.: Markov random fields in image segmentation. Found. Trends Signal Process. **5**(1–2), 7–13 (2012). Author, F.: Article title. Journal **2**(5), 99–110 (2016)

Image Based Classification of Methods-Time Measurement Operations in Assembly Using Recurrent Neuronal Networks

Patrick Rueckert(✉), Katrin Birgy, and Kirsten Tracht

Bremen Institute for Mechanical Engineering (BIME), University of Bremen, Badgasteiner Str. 1, 28359 Bremen, Germany
rueckert@bime.de

Abstract. Image based classification enables the acquisition and transfer of data from manual assembly workstations into a digital environment. Based on the Methods-Time Measurement method, assembly processes are transformed into short, discrete basic operations that are recognised by means of image processing and used as input data for a multilayer neural network. A recurrent neural network algorithm is investigated for its applicability in combination with the sensor data. The five basic MTM operations reaching, grasping, bringing, releasing, and positioning are classified and additional influencing factors, as well as the implementation of an object recognition, are investigated. The following paper addresses the question of the extent to which manual assembly processes can be reliably derived from visual sensor data and classified by machine learning algorithms.

Keywords: Neuronal networks · Assembly · Image processing

1 Introduction

Digitalization methods enable global companies to make production processes more efficient and flexible, allowing them to respond individually to customer demands while maintaining high product quality. Machine learning and imaging processes are methods that can be used for progress recognition in production and assembly. An essential application is the detection of errors during the process to improve the quality of the assembly processes and components as well as productivity. With the assurance of quality during assembly, conclusions can also be drawn about design and manufacturing errors and measures can be implemented retrospectively [1]. One challenge for recording assembly operations is the complex and multi-variant assembly processes, as well as the recording and analysis of the data and the subsequent error output during the process.

A methodology for a structured recording and analysis of data within the assembly is the MTM (Methods-Time Measurement) method [2]. MTM is used to perform a temporal evaluation of manual assembly processes in different levels of detail [3]. In the MTM method, manual assembly movements are divided into basic movements. These basic movements are assigned standard times, which result from the type of movement and its influencing variables [5]. Therefore, the assembly time can be determined more

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 53–62, 2023.
https://doi.org/10.1007/978-3-031-16281-7_6

precisely, the more information is known about the influencing variables [4]. These influencing variables are, for example, the movement length and the size of the part to be assembled [6]. Additional influencing factors that are necessary for the complete recording of the time data of the assembly process are, in addition to the actual assembly time, the component weight, the degree of difficulty of the assembly, and the quality requirements [4].

The totality of all MTM analysis procedures is based on the five MTM-1 basic procedures reaching, grasping, bringing, positioning, and releasing. By combining the five basic movements, it is possible to describe 85% of all activities that occur in manual assembly. Since the manual MTM analysis is a very time-consuming procedure, combined time modules were developed, which are made up of the basic movements. In this way, the analysis effort of complex assembly activities can be reduced. Deuse et al. describe a concept for classifying assembly operations using visual sensor data. The basic MTM-1 movements of the hand-arm system are classified by means of a deep-learning architecture [7]. As Deuse et al. explain, it must be possible to correctly classify assembly operations regardless of the direction of their execution, which is a problem because such a large number of movement combinations cannot be represented in any data set [7].

From the idea to use digital methods for recording and analysing MTM data, the central question of the paper can be derived, to what extent elementary processes of manual assembly can be reliably perceived from sensor data, such as depth imaging cameras. In this context, the question has to be investigated which effective networks and relationships exist between sensor data and assembly operations. In order to be able to differentiate between elementary processes on the basis of analysis and evaluation of sensor data, algorithms for machine learning methods are used and further developed. The question is pursued, which models and machine learning algorithms in combination with image sensor data are suitable for reliable and accurate recognition and classification of elementary assembly operations.

2 Classification of Assembly Operations

From the perspective of image recognition, assembly operations can be categorized as dynamic movements of the hand-arm system. Analyzing comparable, existing implementations for the detection of movements of the hand-arm system, a widely researched field is the classification of dynamic hand gestures. Therefore, machine learning algorithms are used to distinguish different movement patterns of hand gestures.

Molchanov et al. follow a sequential approach to dynamic hand gesture recognition in which a short video sequence is first inserted into a 3D CNN (Convolutional Neural Network), extracting temporal-spatial features. These features serve as input to an RNN (Recurrent Neural Network), which aggregates the transitions of multiple sequences [8]. The approach of Chen et al. uses two different types of skeleton detection. Global-motion features, which are needed for detecting the position of the hand in the whole observation space, and finger-motion features, which are used for feature detection of the position of the fingers [10]. A multimodal approach was presented by Lai and Yanushkevich in 2018. They combine a CNN with an RNN to take advantage of their individual strengths.

Here, depth and skeletal data are used [12]. The architecture presented by Zhao et al. is composed of two parts. The first part, which takes three-dimensional skeletal coordinates as input vector, consists of an RNN. The second part, which takes RGB images as input, consists of a CNN [20].

Table 1. Literature review on machine learning algorithms for hand gesture recognition

Paper	Architecture	Dataset	Input data	Accuracy
Molchanov 2016 [8]	R3DCNN	1532 videos	RGB	83,8%
Smedt 2016 [9]	Skeletal recognition	DHG-14/28	RGB, SD	83%
Chen 2017 [10]	Skeletal recognition	DHG-14/28	RGB, SD	84,68% ± 6,67%
Chen 2019 [11]	Skeletal recognition	DHG-14/28 SHREC 17	RGB, SD	85,75%; 86,55%
Lai 2018 [12]	CNN + LSTM	DHG-14/28	RGB, SD	85,46% ± 5.16%
Bai 2018 [13]	SVM	400 videos	DI	95%
Liao 2019 [14]	CNN	DEVISIGN-D; LR_Dataset	RGB	89,8%; 86,9%
Horain 2017 [15]	LSTM	160 sequences	3DP	95%
Molchanov 2015 [16]	3D-CNN	VIVA	DI	77.5 ± 7.9%
Arif 2019 [17]	3D-CNN + LSTM	UCF-101; HMDB-51	RGB	92,9%; 70,1%
Ouyang 2019 [18]	3D-CNN + LSTM	UCF-101; HMDB-51	RGB	93,4%; 68,9%
Le 2018 [19]	CNN	UCF-101; HMBD-51	RGB	94.6%; 70.5%
Zhao 2017 [20]	3D-CNN	NTU RGB + D	RGB, DI	83.74%
Yan 2018 [21]	ST-GCN	NTU RGB + D	RGB, DI	81.5%
Si 2019 [22]	SR-TSL	NTU RGB + D	RGB, DI	84.8%

Summarizing the state of the art, multimodal approaches provide the best results [Table 1]. However, the improvement over monomodal approaches is usually only one to two percent. The modalities used are usually RGB images (RGB), depth images (DI), skeleton data (SD), and three-dimensional point clouds (3DP). Depending on the application, some of these modalities are more suitable than others. In the field of dynamic gesture recognition, skeleton and depth data are often used. Due to the fact that skeletal data is a numerical representation of different points of the hand, in three-dimensional space, this modality is very informative and easy to process. Therefore, it provides the best single result for many of the research contributions considered. The input of skeletal data is often complemented by the simultaneous input of depth data. Since skeleton

recognition requires an error-free recognition of the hand, these are not considered further. The reason for this is that interaction with objects makes this considerably more difficult. Accordingly, CNN and LSTM architectures offer the best solution approach.

3 Experimental Design and Data Analysis

One of the main properties of a dataset in supervised learning is the type and size of the input with the associated output. In general, the number of classes, in this case, the number of assembly operations to be classified, correlates strongly with the size of the available data set. Therefore, it is advantageous if the number of assembly operations to be classified is as small as possible. Nevertheless, as many assembly operations as possible should be covered. For this reason, it makes sense to use the MTM-1 basic movements of reaching, grasping, bringing, releasing, and positioning to represent the assembly operation. By combining these five assembly operations, it is possible to indirectly recognise more complex assembly operations. Through these combination possibilities, the number of output classes used can be limited to five, which has a direct effect on the minimum size of the data set. In addition, the components are stored and trained in the class object. The components include a cube, a washer, a wheel, a rhombus, and a screw. In this case, the output class is limited to five. In addition, training is carried out according to the distinguishing features of arm movement and location of the arm. For arm movement, there are two output classes, static and dynamic. For the location of the arm, there are four classes, assembly carrier, material supply, assembly carrier to material supply, and material supply to assembly carrier. Training the arm movement and the hand position should provide information about its effect on the recognition of the basic movement.

The experimental set-up consists of an assembly workstation that is observed by a Microsoft Kinect v2.0 camera pointed at the workstation from above. In the experiment, a wooden base plate is set up in the middle of the workstation with six screws on which the components are mounted (See Fig. 1).

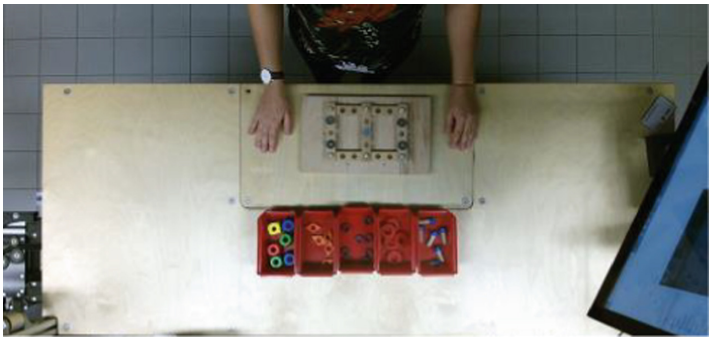


Fig. 1. Workstation for recording the training dataset

3.1 Architecture of the Neuronal Network

The architecture of the neuronal network is based on the work of Rückert et al. (Table 2) [23]. The first layer of the NN is the input layer, which consists of the 53 × 64 pixels of each picture with 3 RGB values. The second layer is a convolutional LTSM layer (ConvLSTM2D) developed by Xingjian Shi. LSTM are special RNN structures that have proven stable and powerful for modeling long-range dependencies [24]. A dropout layer is implemented for reducing overfitting in the neuronal network. It has the same output dimensionality as the layer before. The dense layer (rectified linear unit) is a neural network structure, where all neurons are connected to all inputs and all outputs. After the dense layer, the classification is reduced to 5 dimensions. In order to feed the matrix output of the convolutional and pooling layers into a dense layer for the output, it must first be rolled out (flattened). This is done by a flattening layer, resulting in a single vector with 16960 values. The last dense layer (softmax) creates a 5-dimensional classifier, which resembles the 5 basic MTM-1 movements.

Table 2. Layers of the Neuronal Network

Layer	Rows	Columns	Dimensions	Parameter
Input layer	53	64	3	0
Convolutional layer (ConvLSTM2D)	53	64	64	154624
Dropout	53	64	64	0
Dense (rectified linear unit)	53	64	5	325
Flatten	1	16960	1	0
Dense	1	1	5	84805

3.2 Analysis of the Data Set

A total of 2939 data sequences are recorded and stored. The recorded data set consists of RGB, infrared, and depth image data. The RGB data have a resolution of 480 × 270 pixels, while the infrared and depth image data are recorded and stored with a resolution of 512 × 424 pixels.

Analysing the number of assembly operations in the dataset, each basic operation has the same number of examples. Considering the number of images per assembly operation, positioning has a significantly larger number of images (See Fig. 2). According to the MTM-1 method, each elementary operation takes a different amount of time. Due to that, twice as many images were saved for positioning.

The number of images by objects in the data set shows, that the screw and cube components are used the same number of times since they are used in every run of the assembly scenario, while the wheel, rhombus and washer are used less because they are randomly selected twice in the run of the assembly scenario.

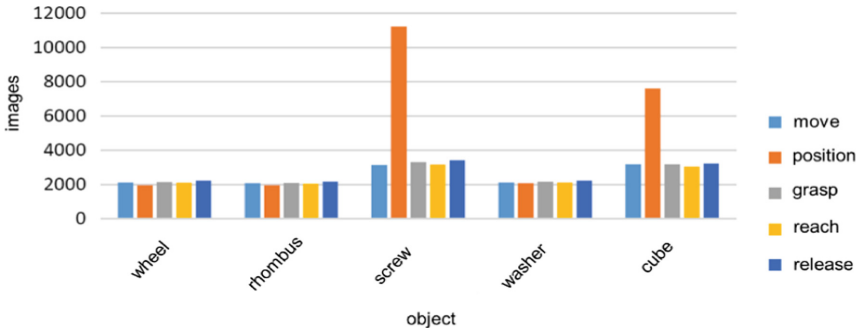


Fig. 2. Number of images by object and assembly operation

3.3 Data Processing

The data processing is based on the results of Rückert and Papenberg [23]. For processing the data before training the algorithm, the image data are extracted from the corresponding folders and read in according to the algorithm developed by Rückert and Papenberg [23].

The data set is further divided into the partial data sets training, validation, and test data. These are divided randomly according to the 60:20:20 principle so that 60% of the data is allocated to the training data set and 20% each to the validation and test data set. This means that the data set is reallocated each time the training is started so that new partial data sets are generated each time. For this reason, there are deviations in accuracy with multiple training runs with the same start parameters. For each run, the distribution of the data set is saved in the respective data files so that the distribution of the partial data sets can be tracked.

In the next step, the infrared images are reduced to a size of 384 × 318 pixels, which is three-quarters of the original size. The reduction in resolution is done to reduce the number of parameters that can be trained. However, this is associated with a loss of information.

The core of the overall architecture used is the ConvLSTM2D architecture This architecture combines the strengths of CNN and LSTM into a common layer within a neural network and is thus significantly better suited for the classification of temporal-spatial data than CNN or LSTM alone.

4 Results

In the test scenarios, the performance of the trained algorithm on the data set is evaluated individually for the classification of assembly operations and objects. For this purpose, the parameter of accuracy is applied. In addition, the individual accuracies are shown in a confusion matrix. The predicted classes are compared with the actual classes. Thus a statement about accuracy can be made. Furthermore, the presented key figures are calculated. With each training of the different classes, different parameters are adjusted to achieve better results. Thereby, runs are made with and without adjustment of the

learning rate, and the variables filter and patience are modified. The results shown are the outcomes for the highest accuracy achieved in the training runs for the classification of the five assembly operations. An accuracy of 87.24% is achieved in the detection of the assembly operation (Fig. 3).

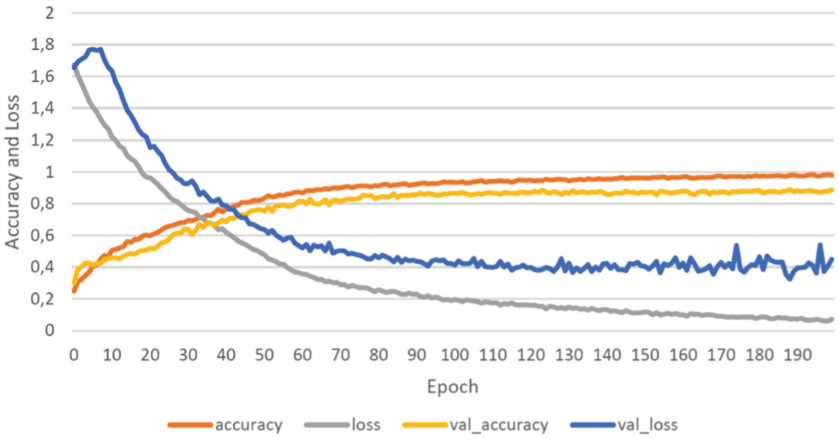


Fig. 3. Accuracy and loss on training of assembly operations

Figure 4 shows the confusion matrix of the training. The main diagonals show a good classification of all assembly operations. Reaching is recognised with 93% and grasping with 96%. Bringing is incorrectly predicted as positioning with 14%. This can be explained by the fact that an object is placed in the hand during bringing and positioning, and that a hand movement can be recognized during both operations. Only 7% of positioning is attributed to bringing and 11% to releasing. The confusion of positioning with releasing corresponds to the expectations since the arm position is above the assembly object for both basic movements and the movement is defined as static. Accordingly, their main difference is the holding or positioning of a component.

prediction

actual class	reach	grasp	bring	position	release
reach	0,93	0,03	0,01	0,02	0,02
grasp	0,00	0,96	0,01	0,03	0,00
bring	0,01	0,00	0,84	0,14	0,02
position	0,01	0,03	0,07	0,79	0,11
release	0,00	0,00	0,05	0,16	0,79

Fig. 4. Confusion matrix for training of assembly operations

During the training runs for the recognition of the objects, not only the parameters filter and patience are changed, but also the sensor. The aim is to test whether a better performance can be achieved when using RGB data in contrast to IR data since the colours

of the components are shown to advantage here. It is possible that misclassification of the cube with the other components occurs, since the cubes are available in several colours, while the other objects only have one colour each (Fig. 5).

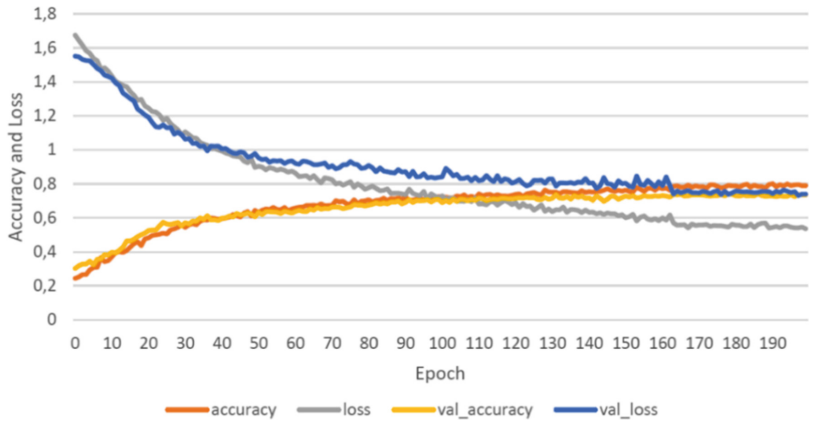


Fig. 5. Accuracy and loss on training of object recognition

The achieved accuracy is 73.3% and is already reached between the 170th and 175th epoch. The adjustment of the learning rate takes place in the 163rd and 197th epoch but has no noticeable influence on the result. However, overfitting begins from the 50th epoch.

When recognising the components, the confusion matrix of the third training run is considered with an accuracy of 73.3%. The cube is mostly categorised correctly and confusion with the rhombus or the screw occurs only in rare cases. The rhombus is correctly categorised in 60% of the cases. However, 18% of the cases are confused with the washer and 11% each with the cube and the wheel. The wheel is recognised well, but it is mistakenly assigned to the washer in 23% of the cases. This can be explained by the round and flat shape of both components, but they differ in diameter. The washer is mistaken for the wheel in only 7% of cases. The washer is well categorised with 79%. The screw has the highest accuracy with 90%. Although no component data was included in the training, contrary to expectations, good results were achieved in object recognition (Fig. 6).

prediction

	cube	rhombus	wheel	washer	screw
cube	0,79	0,08	0,03	0,03	0,07
rhombus	0,11	0,60	0,11	0,18	0,01
wheel	0,10	0,07	0,58	0,23	0,02
washer	0,06	0,08	0,07	0,79	0,00
screw	0,07	0,00	0,02	0,01	0,90

Fig. 6. Confusion matrix for training of object recognition

5 Summary

The detection of the assembly operation achieves an accuracy of 82.56%. The accuracy is close to the outcome of Molchanov [8] (83,8%) and Liao [14] (89,8%, 86,8%), which was also achieved with RGB data as an input. The result is also within the range of the outcome of Arif [17], Ouyang [18] and Le [19] (92,9%–94,5 for less complex and 68,9–70.5% for more complex datasets), matching the complexity of the dataset that lies between the two datasets considered there.

Overall, the implementation of the new training data was successful and it was possible to improve the trained neural network. In future work, the outcome of the object recognition which achieved accuracies of up to 78.06% can be combined with data from the work plan, in order to improve the classification of the assembly operations.

In order to achieve a test result with a higher accuracy as well as a more robust implementation, a data set of sufficient size must be used. Especially with regard to robustness, it is important that as many people as possible are involved in the collection of data, so that individual idiosyncrasies and the handedness of individual people become irrelevant to the predictions of the algorithm. In pre-processing, the images of the individual sequences could be deliberately rotated or important areas could be shifted, which would have a positive effect on the robustness of the system. In addition, a large dataset allows the use of more powerful models without immediate overfitting.

References

1. Henke, J.: Eine Methodik zur Steigerung der Wertschöpfung in der manuellen Montage komplexer Systeme: Frauenhofer Verlag. Dissertation (2015)
2. Bokranz, R., Landau, K.: Handbuch Industrial Engineering. Produktivitätsmanagement mit MTM. 2., überarb. und erw. Aufl. Stuttgart: Schäffer-Poeschel (2012)
3. Syska, A.: Produktionsmanagement Das A - Z wichtiger Methoden und Konzepte für die Produktion von heute. Betriebswirtschaftlicher Verlag Dr. Th. Gabler | GWV Fachverlage GmbH, Wiesbaden, Wiesbaden (2006)
4. Deuse, J., Busch, F.: Zeitwirtschaft in der Montage. In: Lotter, B., Wiendahl, H.-P. (Hrsg.) Montage in der Industriellen Produktion. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29061-9_4
5. Arndt, K.-D.: Arbeitsvorbereitung und Arbeitsplanung. In: Böge, A., Böge, W. (eds.) Handbuch Maschinenbau, pp. 1735–1772. Springer, Wiesbaden (2017). https://doi.org/10.1007/978-3-658-12529-5_110
6. Lotter, B.: Die Primär-Sekundär-Analyse. In: Lotter, B.,Wiendahl, H.-P. (Hrsg.) Montage in der Industriellen Produktion, pp. 49–78. Springer, Heidelberg (2012)
7. Deuse, J., et al. (Hrsg.): Automatic Generation of Methods-Time Measurement Analyses for Assembly Tasks from Motion Capture Data Using Convolutional Neuronal Networks - A Proof of Concept. Advances in Human Factors and Systems Interaction. Springer, Washington, D.C. (2019)
8. Molchanov, P., et al. (Hrsg.): Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
9. De Smedt, Q., et al. (Hrsg.): Skeleton-based dynamic hand gesture recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2016)

10. Chen, X., et al. (Hrsg.): Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition. In: 2017 IEEE International Conference on Image Processing (ICIP) (2017)
11. Chen, X., Wang, G., Guo, H., Zhang, C., Wang, H., Zhang, L.: MFA-Net: motion feature augmented network for dynamic hand gesture recognition from skeletal data. Sensors **19**(2), 239 (2019)
12. Lai, K., Yanushkevich, S.N. (Hrsg.): CNN+RNN depth and skeleton based dynamic hand gesture recognition. In: 2018 24th International Conference on Pattern Recognition (ICPR) (2018)
13. Bai, X., et al. (Hrsg.): Dynamic hand gesture recognition based on depth information. In: 2018 International Conference on Control, Automation and Information Sciences (ICCAIS) (2018)
14. Liao, Y., et al.: Dynamic sign language recognition based on video sequence With BLSTM-3D residual networks. IEEE Access **7**, 38044–38054 (2019)
15. Sarkar, A., Gepperth, A., Handmann, U., Kopinski, T.: Dynamic hand gesture recognition for mobile systems using deep LSTM. In: Horain, P., Achard, C., Mallem, M. (eds.) IHCI 2017. LNCS, vol. 10688, pp. 19–31. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-72038-8_3
16. Molchanov, P., et al. (Hrsg.): Hand gesture recognition with 3D convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2015)
17. Arif, S., et al.: 3D-CNN-based fused feature maps with LSTM applied to action recognition. Future Internet **2019** (2019)
18. Ouyang, X., et al.: A 3D-CNN and LSTM based multi-task learning architecture for action recognition. IEEE Access **7**, 40757–40770 (2019)
19. Wang, L., Zang, J., Zhang, Q., Niu, Z., Hua, G., Zheng, N.: Action recognition by an attention-aware temporal weighted convolutional neural network. Sensors **18**(7), 1979 (2018)
20. Zhao, A., Ali, H., van der Smagt, P. (Hrsg.): Two-stream RNN/CNN for action recognition in 3D videos. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2017)
21. Yan, S., Xoing, Y., Lin, D. (Hrsg.): Spatial Temporal Graph CNNs for Skeleton-Based Action Recognition (2018)
22. Si, C., Jing, Y., Wang, W., Wang, L., Tan, T.: Skeleton-based action recognition with spatial reasoning and temporal stack learning. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11205, pp. 106–121. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_7
23. Rückert, P., Papenberg, B., Tracht, K.: Classification of assembly operations using machine learning algorithms based on visual censor data. In: 8th CIRP Conference of Assembly Technology and Systems, Procedia CIRP, Athens (2020)
24. Shi, X., Chen, Z., Wang, H., et al.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. arXiv (2015)

Pervasive and Ubiquitous Intelligence

FPGA-Based Road Crack Detection Using Deep Learning

Lorenzo Canese, Gian Carlo Cardarilli, Luca Di Nunzio, Rocco Fazzolari, Marco Re, and Sergio Spanò(✉)

University of Rome "Tor Vergata", Rome, Italy
{canese,cardarilli,di.nunzio,fazzolari,re,spano}@ing.uniroma2.it

Abstract. In this paper we propose different FPGA implementations of a Convolutional Neural Network where the use case is to detect road cracks via images. The work is based on the network proposed by the current state of the art in the field. We deploy the network using the MATLAB Deep Learning HDL toolbox on all available AMD-Xilinx platforms, including different data types. In particular, we use the ZC706 and ZCU102 development boards. In order to infer the CNN, we apply a single precision and an 8-bit integer data type quantization. The implementation results show that the detection accuracy of 99.6% is the same of the state of the art, even though the network is quantized. We also obtain a speed-up of the CNN reaching up to 313.2 Frames Per Second while requiring only 45.85 mJ to process one frame. The proposed implementations are therefore a viable solution for a fast and low-power crack detection system.

Keywords: Road cracks detection · FPGA · Deep Learning · Convolutional Neural Networks · Structural Health Monitoring

1 Introduction

Road degradation in the form of cracks can degrade road performance and offer a risk of accidents [1]. Governments all around the world devote billions each year to improve the condition of their road networks [2]. Safety must be prioritized, and authorities are aware of the necessity for appropriate road inspection and maintenance practices. The foundation of road maintenance systems is crack detection, which has become a trending research area in recent years [3]. Manual road crack identification is recognized to be time-consuming, risky, labor-intensive, and subjective. For this reason, automated crack detection technologies have progressively supplanted older approaches in Intelligent Transportation Systems (ITS) [4], as they provide a more reliable and quicker analysis. Moreover, automated technologies can accurately analyze road surface condition and assist the planning and prioritization of infrastructure repair, resulting in better infrastructures with longer service lives [5].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 65–73, 2023.
https://doi.org/10.1007/978-3-031-16281-7_7

In the last years, Machine Learning algorithms found applications in several fields such as communications, security, health etc. [6–9]. This was possible thanks to the increasing availability of computational capabilities and the realization of optimized hardware architectures [10–12]. Various machine learning-based crack detection systems have been presented [13–15], and neural network are extremely likely to replace the older methods.

In our work we propose an FPGA-based Convolutional Neural Network able to efficiently detect road cracks. Our goal is to reach a high throughput while maintaining a low-power consumption.

2 Dataset and Crack Detection Network Architecture

In this work we use the "Concrete Crack Images for Classification" dataset [16] that provides concrete photos. The pictures were gathered from a variety of Middle East Technical University (METU) campus buildings.

The dataset is split in two classes: Negative (without cracks) and Positive (with cracks). An example of the data is shown in Fig. 1.

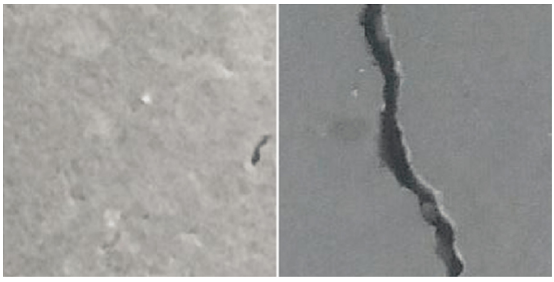


Fig. 1. Two sample images from the dataset: Negative class (no crack) on the left, and Positive class (crack) on the right.

There are 20,000 photos in each class for a total of 40,000 RGB images with a size of 227 by 227 pixels. Zhang's et al. approach [13] was used to create the dataset, consisting in 458 high-resolution photos (4032 × 3024 pixels). Surface quality and lighting conditions widely vary among high-resolution pictures. There is no data augmentation in the form of random rotation or flipping.

Such dataset has been employed in different researches [13–15], beeing Fan's et al. work [15] the actual state of the art in the detection of road cracks. For this reason, we consider the pretrained Convolutional Neural Network (CNN) proposed in their paper, its layers are described in Table 1.

The CNN was designed in MATLAB and it consists of 31 layers. The network performs 7 levels of convolutions followed by operations of batch normalization, ReLU, and Max pooling, with the exception of the last one. This layer, in fact, is followed just by a batch normalization and a ReLu. Finally, a Fully Connected and a softmax layer are used for the actual classification of the pictures.

Table 1. Employed Convolutional Neural Network layers.

Layer	Type	Description
1	Image input	227 × 227 × 3 images with 'zerocenter' normalization
2	Convolution	2 3 × 3 convolutions with stride [1 1] and padding 'same'
3	Batch normalization	Batch normalization
4	ReLU	ReLU
5	Max pooling	2 × 2 max pooling with stride [2 2] and padding [0 0 0 0]
6	Convolution	4 3 × 3 convolutions with stride [1 1] and padding 'same'
7	Batch normalization	Batch normalization
8	ReLU	ReLU
9	Max Pooling	2 × 2 max pooling with stride [2 2] and padding [0 0 0 0]
10	Convolution	8 5 × 5 convolutions with stride [1 1] and padding 'same'
11	Batch normalization	Batch normalization
12	ReLU	ReLU
13	Max pooling	2 × 2 max pooling with stride [2 2] and padding [0 0 0 0]
14	Convolution	16 5 × 5 convolutions with stride [1 1] and padding 'same'
15	Batch normalization	Batch normalization
16	ReLU	ReLU
17	Max pooling	2 × 2 max pooling with stride [2 2] and padding [0 0 0 0]
18	Convolution	32 5 × 5 convolutions with stride [1 1] and padding 'same'
19	Batch normalization	Batch normalization
20	ReLU	ReLU
21	Max Pooling	2 × 2 max pooling with stride [2 2] and padding [0 0 0 0]
22	Convolution	64 11 × 11 convolutions with stride [1 1] and padding 'same'
23	Batch Normalization	Batch normalization
24	ReLU	ReLU
25	Max pooling	2 × 2 max pooling with stride [2 2] and padding [0 0 0 0]
26	Convolution	128 11 × 11 convolutions with stride [1 1] and padding 'same'
27	Batch normalization	Batch normalization
28	ReLU	ReLU
29	Fully connected	2 fully connected layer
30	Softmax	Softmax
31	Classification output	Crossentropyex

3 FPGA Implementation and Deployment

The CNN described in the previous section was coded on FPGA using the MATLAB Deep Learning HDL toolbox [17]. This piece of software allows to deploy a Deep Learning processor on FPGA and to run a pretrained network on it. The processor architecture can be set to manage single precision data (hereafter referred to as "single") or 8-bit integer data (hereafter referred to as "INT8").

At the moment, AMD-Xilinx ZCU102 and ZC706, as well as Intel Arria10SoC development kits, are the only boards compatible with the Deep Learning HDL toolbox (R2022a MATLAB version). We used both AMD-Xilinx ZCU102 and ZC706 in our experiments.

An overview of the FPGA system is shown in Fig. 2. A thorough explanation of the Deep Learning Processor can be found in Span'o et al. work [18].

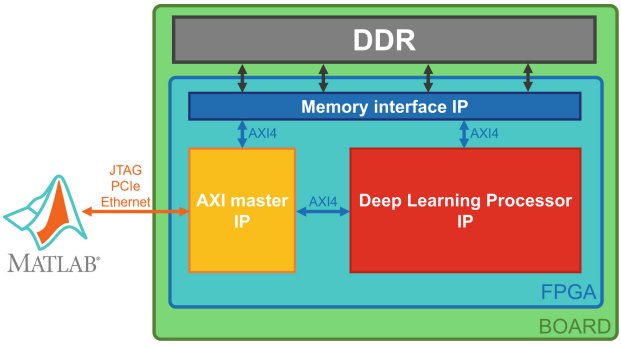


Fig. 2. FPGA-based Deep Learning system overview.

A JTAG, PCIe, or Ethernet interface can be used to communicate with the board. In this work we employed an Ethernet connection. The communication between the computer client and the FPGA is handled by an AXI master module. An AXI4 interface is used for all internal communication in the device.

The Xilinx Memory Interface Intellectual Property (IP) managed the data exchange between the Deep Learning processor and the board's DDR RAM. The weights of the neurons, the kernels of the CNN filters, and, if necessary, the inference images are stored in this memory.

Figure 3 depicts a detailed description of the blocks included within the Deep Learning processor IP.

The system is divided into two parts: one for Convolution (Conv) layer processing and the other for Fully Connected (FC) layer processing. The details of the structure are listed below.

- **Convolution and Fully Connected Processor Controllers.** They are the schedulers of the machine instructions from the client.
- **Memory access Processors.** They are the interface with the Memory Controllers and, hence, with the RAM.

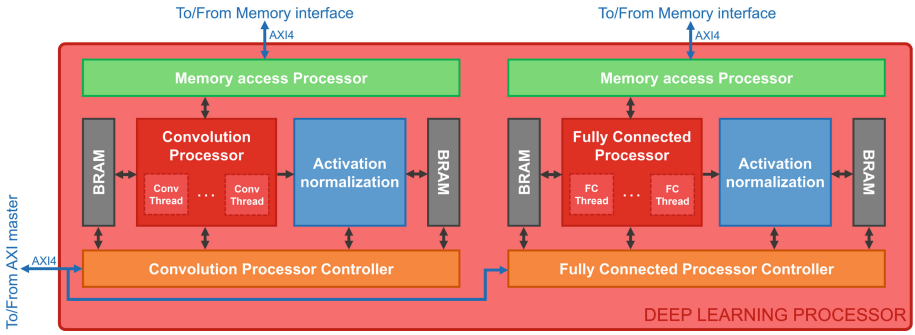


Fig. 3. MATLAB Deep Learning processor architecture.

- **Activation normalization.** They compute all the non-linearities of the networks (i.e. ReLUs) and normalize the data.
- **Block RAMs (BRAM).** They are used as the cache of the Deep Learning processor.
- **Convolution Processor**. It is the core of the convolution computations. It is possible to instantiate up to 64 parallel Threads.
- **Fully Connected Processor**. It is the core of the fully connected (neurons) computations. It is possible to instantiate up to 16 parallel Threads.

3.1 Implementation Results

The Deep Learning HDL toolbox provides a set of standard processors configurations able to deal with single or INT8 data types. In our experiments we considered both types on the ZC706 and ZCU102 platforms, leading to a total of 4 configurations.

The implementation results for all the mentioned setups are presented in Table 2. Please note that the power dissipation was evaluated at the respective "actual" clock frequency.

It is straightforward to see how, on the same device, the single solutions require an increased amount of LUTs and BRAMs. In particular, the number of LUTs used as RAMs (LUTRAMs) is about as twice as the number required on INT8 solutions. INT8 configurations require a double amount of DSPs, this seems to go against the trend. However, such event is easy to explain since FPGA DSPs are designed on purpose to deal with Fixed Point numbers. This means that INT8 solutions are more efficiently mapped on DSPs, thus explaining the increased amount of LUTs that must be used to cover the computational needs of single configurations. Power dissipation is almost the same if a certain board is considered. Of course, Ultrascale + devices such as ZCU102 have more hardware resources and they are more power-efficient than standard devices like ZC706. This matter will be examined in the next section.

Table 2. Implementation results for different boards and data types.

Resource	ZC706 INT8	ZC706 single	ZCU102 INT8	ZCU102 single
LUT	142,028 (64.97%)	146,345 (66.95%)	218,509 (79.72%)	256,476 (93.58%)
LUTRAM	13,600 (19.32%)	25,158 (35.74%)	15,092 (10.48%)	34,203 (23.75%)
FF	154,045 (35.23%)	156,567 (35.81%)	213,125 (38.88%)	256,647 (46.82%)
BRAM	435 (79.82%)	544.5 (99.91%)	453 (49.67%)	581 (63.71%)
DSP	269 (29.89%)	149 (16.56%)	771 (30.6%)	371 (14.72%)
IO	117 (32.32%)	117 (32.32%)	52 (15.85%)	52 (15.85%)
BUFG	8 (25%)	8 (25%)	12 (2.97%)	13 (3.22%)
MMCM	2 (25%)	2 (25%)	2 (50%)	2 (50%)
PLL	1 (12.5%)	1 (12.5%)	1 (12.5%)	1 (12.5%)
Clock frequency	ZC706 INT8	ZC706 single	ZCU102 INT8	ZCU102 single
Actual	90 MHz	100 MHz	250 MHz	220 MHz
Maximum	92.98 MHz	101.58 MHz	250.01 MHz	224.94 MHz
Power dissipation	ZC706 INT8	ZC706 single	ZCU102 INT8	ZCU102 single
Total	5.446 W	5.598 W	14.364 W	13.731 W
Static	0.280 W	0.288 W	0.822 W	0.818 W
Dynamic	5.166 W	5.31 W	13.542 W	12.913 W

4 Performance Analysis

In order to asses the performance of the detection system, we inferred the CNN on the aforementioned devices. The actual boards used for the testing are shown in Fig. 4.

4.1 Detection Accuracy

We took 2,000 pictures from the dataset, 1,000 for each class. Of course, none of such images was used to train the network. We also quantized the original double-precision CNN [15] in a single and INT8 format to be run on the selected FPGAs.

The quantization process did not affect at all the detection performance of the network. In fact, all the versions, including the original one, achieved the same accuracy of 99.6%. The obtained confusion matrix is shown in Fig. 5. Only 4 pictures both for the Negative and Positive class were misclassified.

Obtaining the same detections for every data type means that the quantization process was correctly performed and that the FPGA versions of the CNN can be confidently used to detect the cracks.

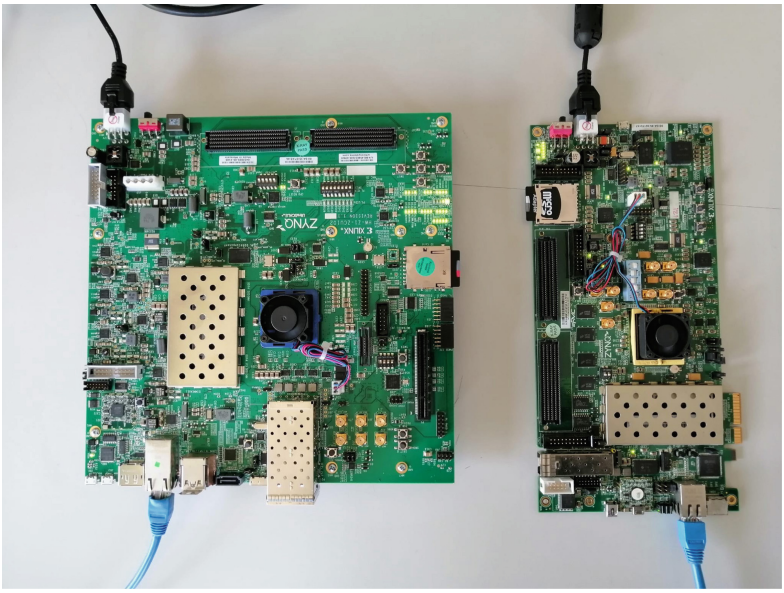


Fig. 4. AMD-Xilinx ZCU102 (left) and ZC706 (right) boards used in the experiments.

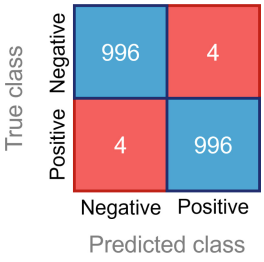


Fig. 5. Confusion matrix of the crack detection network.

4.2 Throughput and Energy Efficiency

In the state of the art [15] the classification was carried out by a Intel Core i7-8700K CPU obtaining a throughput of 208.3 Frames Per Second (FPS). This processor has a typical power consumption of 95 W [19]. Unfortunately, other comparisons are not possible since the considered couple CNN-Dataset has been used only in the above cited work.

Considering the above considerations, the i7 CPU requires 450 mJ to process one frame.

We evaluated the throughput and the energy required to process a frame on the FPGA devices obtaining the results in Table 3.

The best performing configuration both in terms of FPS and energy is the ZCU102 INT8 one. It throughput is more than 100 FPS higher than the CPU implementation and the energy efficiency is one order of magnitude higher (45.86 mJ vs 450 mJ).

Table 3. CNN performance for different boards and data types.

Metric	ZC706 INT8	ZC706 single	ZCU102 INT8	ZCU102 single
Throughput	58.8 FPS	20.8 FPS	313.2 FPS	125 FPS
Energy/Frame	92.62 mJ	269.13 mJ	45.86 mJ	109.85 mJ

We can state that, since the detection accuracy between CPU and FPGA is the same, our proposed implementations are a valid solution for a fast and low-power crack detection system.

5 Conclusion

In this paper we proposed different FPGA implementations of a Convolutional Neural Network where the use case was detect road cracks via images. The work was based on the network proposed by the current state of the art in the field.

We deployed the network using the MATLAB Deep Learning HDL toolbox on all available AMD-Xilinx platforms, including different data types. In particular, we used the ZC706 and ZCU102 development boards. In order to infer the CNN, we applied a single precision and an 8-bit integer data type quantization.

The implementation results showed that the detection accuracy of 99.6% is the same of the state of the art, even though the network was quantized. We also obtained a speed-up of the CNN reaching up to 313.2 Frames Per Second while requiring only 45.85 mJ to process one frame.

The proposed implementations are therefore a viable solution for a fast and low-power crack detection system. As future development, we plan to design a new CNN architecture designed to further improve the throughput.

Acknowledgements. The authors would like to thank Advanced Micro Devices, Inc. (AMD) for providing the FPGA hardware and software tools with the Xilinx University Program.

References

1. Fan, R., Liu, Y., Yang, X., Bocus, M.J., Dahnoun, N., Tancock, S.: Real-time stereo vision for road surface 3-d reconstruction. In: 2018 IEEE International Conference on Imaging Systems and Techniques (IST), pp. 1–6. IEEE (2018)
2. Oliveira, H., Correia, P.L.: Automatic road crack segmentation using entropy and image dynamic thresholding. In: 2009 17th European Signal Processing Conference, pp. 622–626. IEEE (2009)
3. Nguyen, T.S., Avila, M., Begot, S.: Automatic detection and classification of defect on road pavement using anisotropy measure. In: 2009 17th European Signal Processing Conference, pp. 617–621. IEEE (2009)
4. Zhan, J., Dong, S., Hu, W.: IOE-supported smart logistics network communication with optimization and security. Sustain. Energy Technol. Assess. **52**, 102052 (2022)

5. Oliveira, H., Correia, P.L.: Automatic road crack detection and characterization. IEEE Trans. Intell. Transp. Syst. **14**(1), 155–168 (2012)
6. Giuliano, R.: The next generation network in 2030: applications, services, and enabling technologies. In: International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), vol. 2021, pp. 294–298 (2021)
7. Cardarilli, G.C., et al.: An FPGA-based multi-agent reinforcement learning timing synchronizer. Comput. Electric. Eng. **99**, 107749 (2022)
8. Jaber, A.A., Ali, K.M.: Artificial neural network based fault diagnosis of a pulleybelt rotating system. Int. J. Adv. Sci. Eng. Inf. Technol. **9**(2), 544–551 (2019)
9. Giuliano, R., Mazzenga, F., Vizzarri, A.: Satellite-based capillary 5G-mMTC networks for environmental applications. IEEE Aerosp. Electron. Syst. Magaz. **34**(10), 40–48 (2019), cited By: 14
10. Cardarilli, G.C., et al.: A pseudo-softmax function for hardware-based high speed image classification. Sci. Reports **11**(1) (2021), cited By: 3
11. Cardarilli, G.C., et al.: A parallel hardware implementation for 2-d hierarchical clustering based on fuzzy logic. IEEE Trans. Circuits Syst. II: Exp. Briefs **68**(4), 1428–1432 (2021), cited By: 6
12. Sciuto, G.L., Susi, G., Cammarata, G., Capizzi, G.: A spiking neural network based model for anaerobic digestion process. In: 2016 International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM), pp. 996–1003. IEEE (2016)
13. Zhang, L., Yang, F., Zhang, Y.D., Zhu, Y.J.: Road crack detection using deep convolutional neural network. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 3708–3712. IEEE (2016)
14. Ozgenel, C.F., Sorguc, A.G.: Performance comparison of pretrained convolutional neural networks on crack detection in buildings. In: Proceedings of the International Symposium on Automation and Robotics in Construction (ISARC), vol. 35, pp. 1–8. IAARC Publications (2018)
15. Fan, R., et al.: Road crack detection using deep convolutional neural network and adaptive thresholding. In: 2019 IEEE Intelligent Vehicles Symposium (IV), pp. 474–479. IEEE (2019)
16. Caglar, F., Ozgenel, R.: Concrete crack images for classification. Mendeley Data **2** (2019)
17. The MathWorks, Inc.: Deep Learning HDL Toolbox (2022). https://www.mathworks.com/products/deep-learning-hdl.html. Accessed 20 Mar 2022
18. Spano, S., Canese, L., Cardarilli, G.C.: Profiling of CNNS using the MATLAB FPGA based deep learning processor. In: 17th International Conference on PhD Research in Microelectronics and Electronics. IEEE (2022) (in press)
19. Intel Corporation. Intel® Core™ i7-8700K Processor (2022). https://ark.intel.com/content/www/us/en/ark/products/126684/intel-corei78700k-processor-12m-cache-up-to-4-70-ghz.html. Accessed 20 Mar 2022

Simple Non Regressive Informed Machine Learning Model for Prescriptive Maintenance of Track Circuits in a Subway Environment

Andrea Garrone[1], Simone Minisi[2], Luca Oneto[1(✉)], Carlo Dambra[2], Marco Borinato[3], Paolo Sanetti[3], Giulia Vignola[3], Federico Papa[3], Nadia Mazzino[3], and Davide Anguita[1]

[1] University of Genoa, Via Opera Pia 11a, 16145 Genova, Italy
{andrea.garrone,luca.oneto,davide.anguita}@unige.it
[2] ZenaByte s.r.l, Via Cesarea 2, 16121 Genova, Italy
{simone.minisi,carlo.dambra}@zenabyte.com
[3] Hitachi Rail STS, Via Paolo antovani 3/5, 16151 Genova, Italy
{marco.borinato,paolo.sanetti,giulia.vignola}@hitachirail.com,
{federico.papa,nadia.mazzino}@hitachirail.com

Abstract. A Track Circuit (TC) System enables automatic train detection and protection functions. A typical TC failure is the False Occupancy (FO), i.e., a TC results occupied for a certain period of time when there is no train on it. FO can be resolved by manually driving a train on the faulty TC with a waste of time and resources. After a FO, operators, based on their experience, need to decide if there is an actual need for maintenance or if the problem is due to other temporary causes. In this paper, we propose a Simple Informed Machine Learning (ML) based model able to automatically prescribe maintenance after a FO fully leveraging on the operator experience. However, ML models in modern industrial MLOps pipelines demand continuous data collection, model re-training, testing, and monitoring, creating a large technical debt. In fact, one of the main requirements of these pipelines is to not be regressive, i.e., not simply improve average performance but not introduce mistakes previously not present (negative flips). In this work we face this problem by empowering the proposed ML with Non Regressive properties. Results on real data coming from a portion of an Italian Metro managed by Hitachi Rail STS will support our proposal.

Keywords: Informed Machine Learning · Non regressive updates · Prescriptive maintenance · Track circuits

1 Introduction

The Track Circuit (TC) System [15] is a wayside component of the Central Control System that allows the transmission of digital cab signalling data in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 74–83, 2023.
https://doi.org/10.1007/978-3-031-16281-7_8

order to enable automatic train protection functions. A TC System is composed of multiple electrical devices, called TCs, exploited to detect the presence or absence of trains on rail tracks. Despite the robust design of the TCs, failures may still occur due to degradation effects of physical parts (e.g., insulated joints) or hardware-related problems (e.g., defective boards and wrong boards calibration) and many solution have been proposed in the literature [1,2,4,6,12,19].

The first objective of our research is to prescribe whether a maintenance operation needs to be performed during the maintenance night shift based on the occupancy events that occurred during the day (or half day). For this purpose, we will leverage on a Simple Informed Machine Learning (ML) based model [5, 8,14,18].

As we will describe later, our problem is characterized by two main issues: data is extremely scarce and also unbalanced. For these reasons, we needed to keep the approach as simple as possible. In fact, even if the no-free-lunch theorem [20] tells us that the choice of an algorithm strongly depends on the specific application, the general idea behind ML is that the more data you have the more complex the algorithm can be and vice versa [14]. In particular, when the dataset is big and data has a particular natural structure (e.g., graphs, images, and speech) Deep ML is nowadays the most effective approach [7] while for small datasets Shallow ML is still competitive or even the best solution[1] [14]. The difference between the two approaches is simple. Shallow ML models require first to carefully handcraft features (the representation) able to fully and synthetically represent the raw data and then feed these features to classical ML models [14]. Deep ML models instead are able to learn the representation directly from the data, and this is why there is a need for more data [7]. For this reason, in our application we will rely on the simple yet effective Shallow ML model which also easily allow to being informed [8,18] with the domain knowledge and the experience of the operators by careful handcrafting feature able to fully and synthetically represent the operators experience. On top of these features, XGBoost [3], carefully tuned with rigorous statistical procedures [11], will be exploited. Then, methods to mitigate problems related to the unbalance of the dataset will be exploited [9].

Nevertheless, in real, high-stake, and mission-critical applications like the one we are facing in this paper, building the initial ML model is only the beginning of the process. Maintenance of ML models in modern industrial development pipelines (known as MLOps [10]) requires implementing processes for continuous data collection, model updating, and monitoring, thereby creating a large technical debt [13]. New models usually improves the overall performance accuracy (i.e., the average number of error), nevertheless, they can still introduce errors on specific predictions that the previous versions of the models did not make (the so called negative flips) showing a so called regressive behavior [21]. ML models regression can cause their post-processing pipelines to break, requiring specialized maintenance interventions, since modern ML empowered architectures contain several components in addition to ML-based modules.

[1] See the results in Kaggle www.kaggle.com, the most popular Machine Learning competition website, where the XGBoost [3] algorithms is the top winner algorithm.

For this reason, in this paper, we will empower our Simple Informed ML model for Prescriptive Maintenance of TCs with non regressive properties building a Simple Non Regressive Informed ML model.

In order to test the quality of our proposal we will exploit six months of data about a subsection of an important Italian Metro managed by Hitachi Rail STS[2]. The rest of the paper is organized as follows. Section 2 formalizes the problem and describes the available data. Section 3 presents our proposal for a Simple Non Regressive Informed ML model for Prescriptive Maintenance of TCs. Section 4 shows the results of applying the methodology presented in Sect. 3 to solve the problem formalized in Sect. 2 with the data presented in the very same section. Section 5 concludes the paper.

2 Problem Formalization and Available Data

TCs are composed of two parts (Fig. 1): an electronic board that communicates with the Computer-Based Interlocking system to enable train control operations and a physical component that comprises the electrical network together with other equipments such as transmission cables and directional relays. In order to enable train detection functions, each TC is associated with a particular rail track section which is electrically isolated from the others. In particular, insulated joints are exploited to limit current leakage between adjacent sections, whereas impedance bonds are used to both allow direct current to pass from one section to the adjacent one and block alternating current inside the section. This structure allows the current signal to run through the rails to the end of the segment in case there is no train on the track circuit. On the contrary, if a train is on the track, the wheels cause a short circuit, preventing the signal from reaching the receiver.

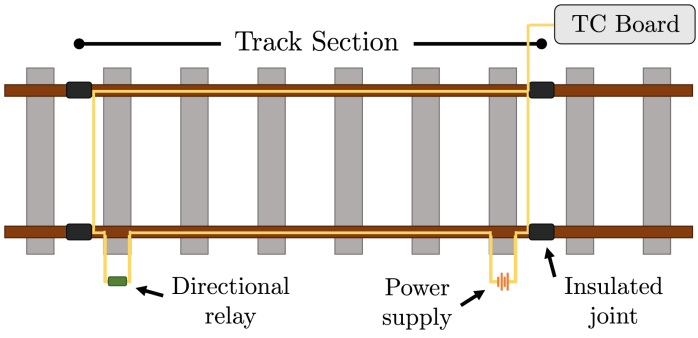


Fig. 1. Track circuit.

[2] All data have been anonymized through the paper (i.e., name of the subsection of the Italian Metro, the number of TCs, the number of faults, etc.) because of confidentiality issues.

The problem we are addressing in this research is related to the TCs False Occupancy (FO) fault. FOs occur when a TC results occupied for a certain period of time when there is no train on it and it may be resolved by manually driving a train onto the faulty track section. This solution introduces delays since the train must be driven by sight but avoids the metro service from being stopped. If this solution does not work a maintenance team needs to be sent as soon as possible. If this solution solves the FO problem, the operators, based on their experience, will decide whether a maintenance intervention will be required at night in order to better understand the root causes of the problem and eventually resetting the TC board or switching the primary board with the backup board. In fact, the FO can be due both to actual problems in the components of the TC Systems and both to ferrous flakes or high levels of humidity.

The task then is to prescribe whether a maintenance operation needs to be performed during the maintenance night shift based on the occupancy events that occurred during the day (or half day, i.e., the first half).

The data exploited in this research is coming from Hitachi Rail STS and refers to a subsection of an important Italian Metro covering six month between 2021 and 2022. Data is coming from two sources: TCs logs and maintenance. TCs logs containing all the events and alarms coming directly from the Central Control System. These logs provide information about alarms related to the Track Circuit system (e.g., board failures and false occupancies), Track circuits occupancy events (e.g., "Track occupied"), trains movements events (e.g., Train 'TrainID' moves from track 'TrackID1' to track 'TrackID2'), and alarms related to other components such as switches and trains doors. A typical Central Control System log record is composed of multiple fields depending on the type of information contained. Maintenance data containing past corrective maintenance interventions with additional information about the resolution procedure.

TCs FOs are not frequent, and consequently a limited number of samples are available. Moreover, most of the FOs do not actually need maintenance interventions and so the resulting dataset is unbalanced, i.e., the percentage of times that we need to prescribe maintenance after an FO is very limited with respect to the times in which operation intervention is not needed.

3 Simple Non Regressive Informed Data Driven Model

The proposed Simple Non Regressive Informed Data Driven Model will be constructed in five steps: (i) Engineering the Features to Inform the ML models, (ii) Choosing a Simple ML model[3], (iii) Empowering the ML model with the ability of effectively handling unbalanced datasets, (iv) Empowering the ML model of Non Regressive properties during models updates[4], and (v) Tuning (the hyperparameters) and assessing the performance of the ML model.

[3] We choose XGBoost after test many different other algorithms that we do not report here because of space constraints.

[4] In fact, every week the model is updated on the system by retraining it with the additional data collected during that week.

Regarding Step (i), starting from the raw logs, in which each record contains an event or alarm related to a specific TC, we first performed some cleaning operations by removing duplicates, filling missing values, correcting inconsistencies, and filtering out all the useless information (i.e., events and alarms that were out of interest for our research based on the experience of the operators). Then we started transforming the data in order to obtain a structure that was better suitable for our analysis. This phase required multiple aggregation steps guided by the operators hints on the problem. We first aggregated alarms and events records related to the same TC and to the same sequence of occupancy. In this way, we obtained a dataset in which each record contains a series of Boolean values describing all the possible events and alarms that can occur between an occupancy and the following one (i.e., '1' if the event/alarm has occurred, '0' otherwise) together with some additional information (i.e., time in which the TC resulted occupied, unoccupied, sequentially occupied, and sequentially unoccupied). Once the data has been transformed from single events and alarms into sequences of occupancies per TC, we separated them into two different datasets: one containing healthy sequences (i.e., sequence with no meaningful/relevant events/alarms) and one containing unhealthy sequences (i.e., sequences containing at least one meaningful/relevant events/alarms). Each of the datasets was then aggregated per day and TC. This operation required the computation of new aggregated values: the features describe [5] the series of time slots in which the TC resulted occupied, unoccupied, sequentially occupied, and sequentially unoccupied for each TC in a day and the number of occurrences of each event/alarm in the sequences on a day. Then we removed records of those days in which a specific TC had registered no unhealthy sequences. After performing this operation we managed to have the healthy sequence dataset having the same length as the unhealthy sequence one. The final step consisted in merging the two datasets by Full day, namely data up to 10:00 p.m. (or by Half Day, namely data up to 5:00 p.m.), and TC obtaining a dataset in which the feature vector has the structure reported in Table 1 and the target is whether an actual maintenance was needed or not.

Regarding Step (ii), one can observe that the datasets (Full and Half Day) created at Step (i) are typical datasets for ML regression. After testing several Shallow and Deep ML models we decided to opt for the XGBoost [3] because of its simplicity and effectiveness. XGBoost is characterized by several hyperparameters, the most important ones are the L2 regularization hyperparameters λ_2 (default value 1) the learning rate of the gradient η (default value 0.3), the max dept of arch tree d (default value 6), the minimum loss reduction γ (default value 0), fraction of training to randomly sample from the whole training set for each tree creation f_s (default value 1), and the fraction of feature to randomly sample from the whole featured during each node of each tree creation f_f (default value 1).

Regarding Step (iii), what we noticed is that XGBoost tended to make more mistakes on the minority class of the datasets (Full and Half Day) created at Step (i). In fact, unfortunately for a data scientist but luckily for the reliability

Table 1. Feature vector defined in the feature engineering phase.

Day of the week
N° of "Track occupied" (for unhealthy and healthy)
N° of "Track unoccupied" (for unhealthy and healthy)
N° of "Track sequentially occupied" (for unhealthy and healthy)
N° of "Track sequentially unoccupied" (for unhealthy and healthy)
N° of "False occupancy" (for unhealthy)
N° of "Alarm ID" (for unhealthy and for each meaningful Alarm ID)
Features described in [5] for the series of (sequentially) Occupied/Unoccupied time slots (for unhealthy and healthy)

of the TCs, most of the FOs do not actually require to prescribe an maintenance interventions resulting in a strongly unbalanced dataset. To address this issue we rely on the combination of a general techniques for handling unbalanced datasets, namely undersampling [9] (since SMOTE in this case tends to generate not plausible samples), plugged inside the XGBoost sampling procedure performed during the creating of each tree composing the ensemble [3]. Moreover, we tuned the threshold in the probability output t (default value .5) before making a decision.

Regarding Step (iv), every week the model is updated by retraining it with the new data collected during that week[5]. During model update we need to be sure that the model does not actually regress, namely introduces new errors that previously were not there. We drew inspiration from the work of [21] and we added a constraint to the learning phase of XGBoost. In particular, let us define f_j the old model built at week j and f_{j+1} the new model built at week $j+1$. Let us also define as $\mathcal{D}_j = \{(X_1, Y_1), \cdots, (X_{n_j}, Y_{n_j})\}$ as the dataset built at week j where X is the feature vector (for Full or Half day) described at Step (i) and $Y \in \{\pm 1\}$ indicates the fact that the maintenance does $(+1)$ or does not (-1) need to be prescribed. Note that $\mathcal{D}_j \subseteq \mathcal{D}_{j+1}$. What we want is that for f_{j+1}

$$\sum_{(X,Y)\in\mathcal{D}_{j+1}:f_j(X)Y\leq 0}[f_{j+1}(X)Y > 0] = 0, \tag{1}$$

namely to not introduce with f_{j+1} errors that f_j did not make. Unfortunately, Constraint (1) is non convex and for this reason we will relax it as follows [17]

$$\sum_{(X,Y)\in\mathcal{D}_{j+1}:f_j(X)Y\leq 0} \max[0, 1 - Y f_{j+1}(X)] = 0, \tag{2}$$

which is a convex approximation of Constraint (1). Since plugging this constraint in the XGBoost training phase is not trivial we will rely on the Tikhonov principle [16] adding to the classical XGBoost objective the term

[5] Obviously, in our case, we simulated with our six month of data this process. We started from the end of month 2, in order to have enough samples, and then we retrained every week from the third month on.

$$\lambda_{nr} \sum_{(X,Y)\in\mathcal{D}_{j+1}: f_j(X)Y\leq 0} \max[0, 1 - Y f_{j+1}(X)], \tag{3}$$

where λ_{nr} is a large enough constant (in our case 1000).

Finally, Step (v) is devoted to tune and assess the performance of the final model we proposed. For this purpose we rely on a resampling procedure adapted to the fact that data are "sorted" in time and that every week the model is updated [11]. In particular, at week j we train the model f_j with the data $\mathcal{D}_j$ and we tested it (assess its performance) with $\mathcal{D}_{j+1} \setminus \mathcal{D}_j$. We measured the quality of the model according to four metrics: percentage of errors on positively labeled samples (False Positive %), percentage of errors on negatively labeled samples (False Negative %), percentage of negative flips (i.e., errors made by f_j not made by f_{j-1}) on positively labeled samples (Negative Flips on Positive %), and percentage of negative flips on negatively labeled samples (Negative Flips on Negative %). In order to tune the performance (find the optima hyperparameters) of XGBoost for f_j we perform classical leave-one-out [11] using $\mathcal{D}_j$. During this phase we performed a grid search for $\lambda_2 \in \{0.001, 0.01, 0.1, 1, 10, 100\}$, $\eta \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$, $d \in \{3, 5, 10\}$, $\gamma \in \{0, 0.01, 0.02, 0.04, 0.1\}$, $f_s \in \{0.6, 0.8, 1\}$, $f_f \in \{0.1, 0.2, 0.4, 1\}$, and $t \in \{.2, .4, .5, .6, .8\}$.

4 Experimental Results

In this section we will show the results of applying the methodology presented in Sect. 3 to solve the problem formalized in Sect. 2 with real data coming from a portion of an Italian Metro managed by Hitachi Rail STS.

In particular, in Table 2 we reported the mean and the standard deviation of the metrics defined in Sect. 3 (False Positive %, False Negative %, Negative Flips on Positive %, and Negative Flips on Negative %) for different models. First we reported the vanilla XGBoost with default values (XGBoost), then we show what happens by tuning the hyperparameters XGBoost maximizing the balanced accuracy (+Tuning), then we show what happens by handling the fact that classes are unbalanced (+Unbal), and finally we show what happens by forcing the model to be non regressive (+NoRegr). For details about XGBoost, +Tuning, +Unbal, and +NoRegr please refer to Sect. 3. We reported the results for the two models we developed in this work: the one that uses the events that occurred during the Full Day to prescribe whether a maintenance operation needs to be performed during the maintenance night shift and the one that uses the events that occurred during Half Day.

From Table 2 we can make some observations that hold from both the model that exploits events that occurred during the Full Day and the one that used just the event that occurred in Half Day. Vanilla XGBoost is not effective enough for a real application because of the high number of False Negatives (i.e., necessary maintenance not prescribed). The problem is a bit mitigated by the +Tuning model but yet not practical. The +Unbal model actually addressed the issue strongly reducing the False Negatives with a slight increase in False Positives (which are less of a problem, in practice it is better to make one more unuseful maintenance that missing a required maintenance). Nevertheless, XGBoost,

Table 2. Results on real data coming from a portion of an Italian Metro managed by Hitachi Rail STS for the different models developed in the paper (XGBoost, +Tuning, +Unbal, and +NoRegr) measured with different performance metrics (False Positive %, False Negative %, Negative Flips on Positive %, and Negative Flips on Negative %) based on events that occurred during the Full Day (or Half Day) to prescribe whether a maintenance operation need to be performed during the maintenance night shift.

	XGBoost	+Tuning	+Unbal	+NoRegr
	Half Day			
False Positive %	78.7 ± 20.1	29.6 ± 10.1	6.0 ± 1.3	4.3 ± 1.2
False Negative %	10.8 ± 342	1.2 ± 1.1	3.0 ± 0.8	2.2 ± 0.6
Negative Flips on Positive %	35.3 ± 10.5	16.3 ± 5.6	5.4 ± 1.6	0.9 ± 0.2
Negative Flips on Negative %	2.6 ± 0.7	0.6 ± 0.6	2.5 ± 0.7	0.8 ± 0.2
	Full Day			
False Positive	67.5 ± 18.8	23.4 ± 9.8	4.5 ± 1.3	2.1 ± 0.6
False Negative	8.9 ± 2.5	1.1 ± 0.9	2.2 ± 0.6	1.9 ± 0.4
Negative Flips on Positive %	29.6 ± 8.9	13.7 ± 3.7	3.5 ± 0.8	0.8 ± 0.2
Negative Flips on Negative %	1.7 ± 0.4	0.5 ± 0.5	2.8 ± 0.6	0.5 ± 0.1

+Tuning, and +Unbal exhibit high levels of regressivity which is bad for modern ML empowered software pipelines. The +NoRegr version of the model strongly reduces this phenomena and, moreover and not surprisingly, tends to increase the accuracy of the final model. These results are in agreement with what is discussed in Sect. 3. Finally note that the model that exploits events that occurred during the day is not so much more effective with respect to the one that used the events of half day. This is a quite important result since from half a day we are already able to plan and organize the maintenance teams for the night shifts.

Based on the reported results we can state that the proposed Simple Non Regressive Informed Machine Learning Model is actually effectively able to prescribe maintenance on the Track Circuits in a Subway Environment.

5 Conclusions

In this work we have focused our attention on the prescriptive maintenance of the Track Circuits, devices that enable automatic train detection and protection functions. In particular, we focused on a particular Track Circuit failure, i.e., the False Occupancy, where TC results occupied for a certain period of time when there is no train on it. After a False Occupancy, operators, based on their experience, need to decide if there is an actual need for maintenance or if the problem is due to other temporary causes. In this work we focused on developing an automatic intelligent tool able to prescribe or not the maintenance after a False Occupancy. For this purpose, we propose a Simple Informed Machine Learning based model, namely a data-driven model able to both exploit the physical knowledge about the phenomena provided by the operators (via advanced

feature engineering from the raw logs and data) and historical data. However, building our Simple Informed Machine Learning models is just the first step of our contribution. In fact our model is exploited in a modern industrial MLOps pipeline and demands continuous data collection, model re-training, testing, and monitoring that results in a large technical debt. An update in the model cannot be regressive, i.e., we do not have to simply improve average performance but also not introduce mistakes previously not present (negative flips). For this purpose, we empowered the proposed Simple Informed Machine Learning with Non Regressive properties with a simple yet effective approach. Results on real data coming from a portion of an Italian Metro managed by Hitachi Rail STS showed the effectiveness of our proposal in prescribing maintenance of Track Circuits after False Occupancy in a non regressive way.

Acknowledgements. This research received funding from Hitachi Rail STS through the RAIDLab (https://raidlab.unige.it/) (Railway Artificial Intelligence and Data Analysis Laboratory), a joint laboratory between Hitachi Rail STS and University of Genoa. Moreover this research received funding from the Shift2Rail Joint Undertaking (JU) under grant agreements No. 881574 (IN2SMART2) and No. 101008913 (DAYDREAMS). The JU receives support from the European Union's Horizon 2020 research and innovation programme and the Shift2Rail JU members other than the Union.

References

1. Alvarenga, T.A., Cerqueira, A.S., Filho, L.M., Nobrega, R.A., Honorio, L.M., Veloso, H.: Identification and localization of track circuit false occupancy failures based on frequency domain reflectometry. Sensors **20**(24), 7259 (2020)
2. Chen, J., Roberts, C., Weston, P.: Fault detection and diagnosis for railway track circuits using neuro-fuzzy systems. Control. Eng. Pract. **16**(5), 585–596 (2008)
3. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
4. Consilvio, A., et al.: Prescriptive maintenance of railway infrastructure: from data analytics to decision support. In: International Conference on Models and Technologies for Intelligent Transportation Systems (2019)
5. D'Amato, V., Oneto, L., Camurri, A., Anguita, D.: Keep it simple: handcrafting feature and tuning random forests and Xgboost to face the affective movement recognition challenge 2021. In: International Conference on Affective Computing & Intelligent Interaction (2021)
6. De Bruin, T., Verbert, K., Babuška, R.: Railway track circuit fault diagnosis using recurrent neural networks. IEEE Trans. Neural Netw. Learn. Syst. **28**(3), 523–533 (2016)
7. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
8. Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L.: Physics-informed machine learning. Nat. Rev. Phys. **3**(6), 422–440 (2021)
9. Kaur, H., Pannu, H.S., Malhi, A.K.: A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput. Surv. **52**(4), 1–36 (2019)

10. Mäkinen, S., Skogström, H., Laaksonen, E., Mikkonen, T.: Who needs MLOps: what data scientists seek to accomplish and how can MLOps help? In: IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (2021)
11. Oneto, L.: Model Selection and Error Estimation in a Nutshell. MOST, vol. 15. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-24359-3_9
12. Oukhellou, L., Debiolles, A., Denœux, T., Aknin, P.: Fault diagnosis in railway track circuits using Dempster-Shafer classifier fusion. Eng. Appl. Artif. Intell. **23**(1), 117–128 (2010)
13. Sculley, D., et al.: Hidden technical debt in machine learning systems. In: Neural Information Processing Systems (2015)
14. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, Cambridge (2014)
15. Surhone, L.M., Timpledon, M.T., Marseken, S.F.: Track Circuit: Train, Rail Tracks, Locomotive, Rolling Stock, Fail-safe, Railway Signalling, Relay, Polarity. Betascript Publishing (2010)
16. Tikhonov, A.N., Arsenin, V.Y.: Methods for Solving Ill-Posed Problems. Nauka, Moscow (1979)
17. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
18. Von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., et al.: Informed machine learning-a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Trans. Knowl. Data Eng. (2021)
19. Wang, X., Guo, J., Jiang, L., Fu, J., Li, B.: Intelligent fault diagnosis and prediction technologies for condition based maintenance of track circuit. In: IEEE International Conference on Intelligent Rail Transportation (ICIRT) (2016)
20. Wolpert, D.H.: The supervised learning no-free-lunch theorems. In: Roy, R., Köppen, M., Ovaska, S., Furuhashi, T., Hoffmann, F. (eds.) Soft Computing and Industry. Springer, London (2002). https://doi.org/10.1007/978-1-4471-0123-9_3
21. Yan, S., et al.: Positive-congruent training: towards regression-free model updates. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)

Embedded Implementation of an Algorithm for Online Inertia Estimation in Power Grids

Alessandro Ravera, Valentina Baruzzi, Matteo Lodi, Alberto Oliveri(✉), and Marco Storace

DITEN, University of Genoa, Via Opera Pia 11a, Genoa 16145, Italy
alberto.oliveri@unige.it

Abstract. The energy transition is an issue of major importance worldwide and entails the gradual replacement of fossil fuels technologies with renewable energy sources (RES) for electric power production. However, integrating photovoltaic and wind power plants in traditional power grids threatens the stability of the system if no additional synthetic inertia is provided by control systems. Due to the intermittent nature of RES, the inertia of the power plants and of the entire grid is time-varying, calling the need for online monitoring methods. In this paper, we implement on a microcontroller an algorithm for online estimation of the inertia constant and damping coefficient of individual energy sources. The behavior of this embedded implementation is analyzed with respect to some key parameters and tested on the IEEE 14-bus power system.

Keywords: Renewable energy sources · Inertia estimation · Microcontroller

1 Introduction

Each synchronous generator in a power grid is characterized by its inertia constant H, which quantifies its ability to counteract frequency changes induced by external disturbances, and therefore contributes to system stability [1]. When a power imbalance occurs in the grid, indeed, the electrical frequency deviates from its steady-state value, and large deviations may lead to instability, and then faults. In a traditional power system composed only of synchronous generators, the inertia comes directly from the kinetic energy stored in the rotating masses, as the mechanical rotor speed is directly coupled with the frequency of the generated electrical power. In this condition, the inertia constant of the power grid remains generally steady over the long term.

The presence of renewable energy sources (RES), like photovoltaic and wind power plants, is increasing in power grids, due to environmental policies and the fact that fossil fuel reserves are limited. RES are interfaced to the grid through power converters, with no intrinsic inertia, which may lead to stability issues in the power grid. To face this problem, the controller of the inverter connecting the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 84–94, 2023.
https://doi.org/10.1007/978-3-031-16281-7_9

RES to the grid can be designed to provide a power contribution proportional to the rate of change of frequency, thus emulating an inertial behavior (synthetic inertia), with the support of storage systems. Typically, embedded digital devices (microcontrollers) are exploited to regulate the inverter. In this scenario, due to the intermittent nature of RES and the limited energy and power which can be supplied by storage systems, the resulting global inertia is time-varying.

This justifies the need for methods able to continuously estimate the inertia constant of the system during normal operating conditions, over time scales comparable to meteorological condition changes, without relying on post-fault data [2–4], injection of probing signals [5] or historical data [6,7], that are often not easily accessible. Such algorithms rely on the data measured through phasor measurement units (PMUs), exploiting the information contained in the small power and frequency fluctuations naturally present at steady state in the power grid. In [8] a linear model relating the active power variations to the frequency variations is estimated through the n4sid algorithm [9] to track the inertia constant of the generators and the network. In [10] this method is analyzed and tested on a power grid with RES, introducing improvements such as a shorter delay in detecting inertia constant changes. In [11] the total inertia of a network with varying percentage of RES penetration is monitored with an algorithm for model parameter estimation based on the recursive least-squares approach, achieving an even shorter delay.

When considering individual RES connected to the power grid, equipped with controllers for synthetic inertia generation, measurement delays and noise could significantly affect the real contribution of synthetic inertia. For this reason, it is important to monitor the inertia that is actually provided by the controller, through an inertia estimation algorithm implemented on an embedded device, directly connected to the inverter controller, without the need to send data to a remote computer. The latency of the implemented algorithm should be lower than about 40 ms, which is the typical update time of commercial frequency transducers (for a mains frequency 50 Hz), in order to update the estimation at every frequency measurement. In this work, a microcontroller implementation of a modified version of the algorithm described in [8] and improved in [10] is proposed and discussed, that exploits the *mesh adaptive direct search* (MADS) optimization algorithm. Simulations are performed on the standard IEEE 14-bus power system, to assess the latency and performance of the embedded device with respect to some parameters of both the inertia estimation and the MADS algorithms.

2 Inertia Estimation Algorithm

The well-known swing Eq. [12] describes the dynamics of a synchronous generator. It relates the rotor speed ω, the active electrical power P_e and the mechanical power P_m all measured in *per units* (p.u.):

$$2H\dot{\omega} = P_m - P_e - D\,(\omega - \omega_0) \tag{1}$$

where ω_0 is the normalized rated rotor speed, H [s] is the inertia constant and D (dimensionless) is the damping coefficient. When the power grid works under normal operating conditions the frequencies and powers in the grid are not perfectly

constant, but vary around their steady-state values, exhibiting small oscillations due to fluctuations in the power absorbed by the loads. If turbine governors are not activated we can assume that P_m remains constant, whereas the rotor speed variations $\Delta\omega$ and the electrical power variations ΔP_e can be approximated by the electrical frequency variations Δf and the active power variations ΔP measured through PMUs at the generator bus [8]. We can therefore rewrite Eq. (1) as:

$$2H\dot{\Delta}f = \Delta P - D\Delta f \tag{2}$$

Since the model structure is known, an optimization can be carried out to determine the parameters H and D that best fit the time evolution of the output Δf, given the input profile ΔP. The optimization process considers W samples of the signals Δf and ΔP, with sampling interval Δt, within a moving time window of $W\Delta t$ seconds. The window is shifted of 1 sample (Δt seconds) at each iteration of the algorithm, so that the inertia constant H and the damping coefficient D are estimated every Δt.[1] In this framework, the values estimated at a certain instant refer to the previous W samples.

The cost function is defined as

$$g(H, D; \{\Delta P_i\}, \{\Delta f_i\}) = \sum_{i=1}^{W} \left[\Delta f_i - \hat{\Delta} f_i(H, D)\right]^2 \tag{3}$$

where $\{\Delta P_i\}$ and $\{\Delta f_i\}$ are arrays whose elements ΔP_i and Δf_i represent samples of the electrical power and frequency variations, respectively, within a time window. Each term $\hat{\Delta} f_i(H, D)$ represents the frequency variation estimated by integrating the swing equation with forward Euler method and parameters H and D:

$$\hat{\Delta} f_i(H, D) = \hat{\Delta} f_{i-1}(H, D) - \frac{\tau}{2H}[\Delta P_i + D\hat{\Delta} f_{i-1}(H, D)] \tag{4}$$

with $\hat{\Delta} f_0(H, D) = \Delta f_0$.

Therefore, we estimate H and D by solving a nonlinear constrained optimization problem:

$$\min_{H,D} g(H, D; \{\Delta P_i\}, \{\Delta f_i\}) \tag{5}$$

subject to $H_{min} \leq H \leq H_{max}$ and $D_{min} \leq D \leq D_{max}$.

2.1 Mesh Adaptive Direct Search

To efficiently estimate the inertia and damping with a microcontroller, a suitable algorithm must be chosen to solve problem (5). The *mesh adaptive direct search* (MADS) algorithm [13] represents a good candidate, as it does not require evaluating derivatives of the cost function and only needs computing the cost function g at some poll points. A microcontroller implementation of the MADS algorithm to solve a nonlinear control problem has been recently proposed in [14]. The algorithm is briefly summarized in the following. By defining the parameter vector

[1] A different update time, greater than Δt, could also be used.

$z = [H, D]^T \in \mathbb{R}^2$ and the domain $\Omega = \{z \in \mathbb{R}^2 : z_{min} \leq z \leq z_{max}\}$, where the inequalities should be considered element-wise, problem (5) can be recast as

$$\min_{z \in \Omega} g(z). \tag{6}$$

For a generic parameter vector $z \in \mathbb{R}^n$ MADS is an iterative algorithm which, starting from an initial point z_0, at each iteration evaluates the cost function g in $2n$ trial points $\hat{z}_k^j$ $(j = 1, \ldots, 2n)$, to get closer to a local minimum of g. At the end of k-th iteration, a new point z_{k+1} is selected with a certain strategy, which will be explained later. An important requirement for the algorithm convergence is that the trial points $\hat{z}_k^j$ explored during an iteration belong to a *mesh*, generally composed of points lying on a regular grid in $\mathbb{R}^n$, whose coarseness is determined by a parameter δ_k (see Fig. 1). Each iteration is composed of an optional *search step* (skipped in our application) and a mandatory *poll step*. The poll step consists of a local exploration around the current point z_k. In particular, $2n$ trial points $\hat{z}_k^j$ are generated within a *frame* (red square in Fig. 1) with size $\Delta_k \geq \delta_k$, i.e., such that $||\hat{z}_k^j - z_k||_\infty \leq \Delta_k$ for $j = 1, \ldots, 2n$. A typical choice (adopted in this work) is $\delta_k = \min\{\Delta_k, (\Delta_k)^2\}$.

In our application, it is always possible to choose the initial point $z_0 \in \Omega$, then an *extreme barrier* is used, meaning that any trial point $\hat{z}_k^j \notin \Omega$ is discarded. As soon as a trial point $\hat{z}_k^* \in \Omega$ is found such that $g(\hat{z}_k^*) < g(z_k)$, the iteration is declared successful. Otherwise, if none of the trial points contained in Ω leads to a reduction of the cost function, the iteration is declared unsuccessful. If the iteration is successful, the frame size (and therefore the mesh size) is increased, in particular $\Delta_{k+1} = \tau^{-1}\Delta_k$, with $\tau \in (0,\ 1)$ and the next point is $z_{k+1} = \hat{z}_k^*$. If the iteration is unsuccessful, the frame size (and therefore the mesh size) is decreased, in particular $\Delta_{k+1} = \tau\Delta_k$, and $z_{k+1} = z_k$. The algorithm ends when the number of iterations reaches a maximum value K, i.e., when $k > K$. The algorithm returns the optimal point $z^* = z_{K+1}$.

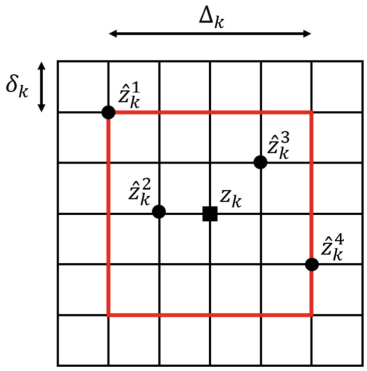


Fig. 1. Illustration of a generic iteration k. The mesh points are the intersections of the black lines. The frame is identified by the red thick square. The solid black square indicates the current point z_k. Black dots indicate the poll points $\hat{z}_k^j$.

The tuning parameters of the algorithm are the maximum number of iterations K, the initial frame size Δ_0, and the coefficient τ. Δ_0 and τ determine the convergence speed of the algorithm, allowing for larger or smaller "jumps" in the search of the optimum.

3 Results

The well known IEEE 14-bus power system [15], shown in Fig. 2, with stochastic loads is used as a case study. Only synchronous generators (G1, G2, G3, G6, G8) are considered here, as the aim of this work is to validate the performance of the embedded implementation of the inertia estimation algorithm; an example with RES can be found in [10]. Simulations of the considered power network are performed through the circuit simulator PAN [16]. The real inertia constants H^{NOM} and rated powers P_0 of the generators in the IEEE 14-bus grid are listed in Table 1. The damping coefficient is equal to 2 for all generators. The inertia of G1 is halved after 250 s, whereas the inertia of G3 is doubled after 500 s. A detailed description of the simulation framework can be found in [10].

We implemented the inertia estimation algorithm on a (low-cost) STM32F4 microcontroller, with a clock frequency of 168 MHz. We set $W = 10$, $H_{min} = 0.1$ s, $H_{max} = 50$ s, $D_{min} = 0$, and $D_{max} = 50$, whereas several values of K, Δ_0 and τ have been used to evaluate the performances of the algorithm. Every $\Delta t = 0.5$ s, samples of Δf and ΔP are sent to the microcontroller through the

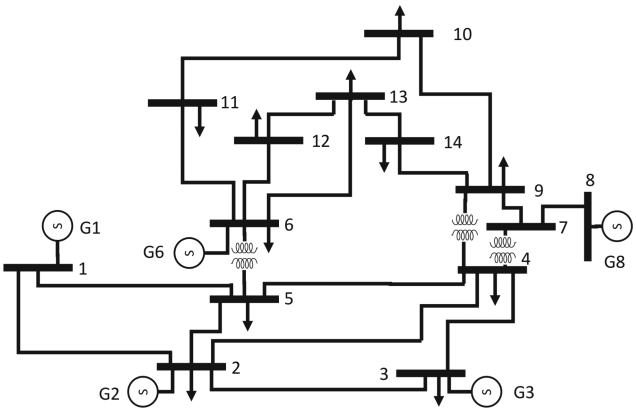


Fig. 2. The IEEE 14-bus power system

Table 1. Inertia H^{NOM} and rated power P_0 of all generators in the IEEE 14-bus grid.

	G1	G2	G3	G6	G8
H^{NOM} [s]	5.148	6.540	6.540	5.060	5.060
P_0 [MW]	615	60	60	25	25

RS232 interface. After a latency time, the estimated values of H and D are sent back to the PC through the same interface. The latency of the algorithm is not constant because each poll step ends as soon as a better point is found. The maximum latency L, however, depends on both K and W, as each algorithm iteration evaluates Eq. (3). From the algorithm operation flow we deduced the following relationship between L and K, W:

$$L = (c_1 W + c_2)K + c_3 \tag{7}$$

Measurements of the maximum latency resulting from different values of W and K allowed fitting this model to data through least squares optimization, thus obtaining coefficients $c_1 = 0.06115$ ms, $c_2 = -0.05713$ ms, and $c_3 = 0.1859$ ms. The fitting error is less than 3%. Figure 3 shows the boundary of the region, on the plane (W, K), where the maximum latency L of the embedded device is lower than 40 ms, namely, the region where the estimation can be updated at every frequency measurement. This region (below the red boundary) would enlarge if a higher-performance device were used.

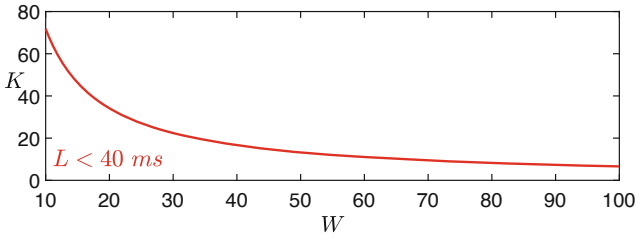


Fig. 3. The red curve is the boundary of the lower region, where the embedded implementation has a maximum latency below 40 ms. (Color figure online)

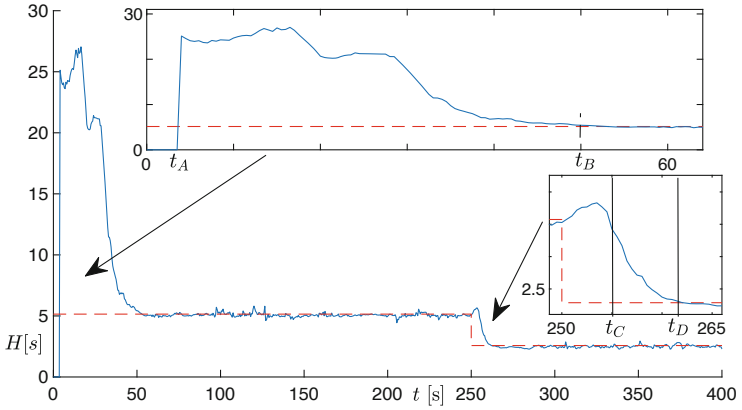


Fig. 4. Time evolution of the estimated inertia of G1, obtained with $K = 10$, $\tau = 0.7$, and $\Delta_0 = 0.01$. The red dashed curve corresponds to the nominal value.

The initial condition for MADS at the first time instant is $H = 25$ s and $D = 25$ (the point in the center of the domain). For the next time instants, the initial condition is set as the optimum found in the previous instant. Figure 4 shows (blue curve) a typical profile of H obtained by the algorithm implemented on the microcontroller with $K = 10$, $\tau = 0.7$, and $\Delta_0 = 0.01$, compared to the real inertia H^{NOM} of generator 1 (red curve).

Notice that, for $t < t_A$, the algorithm returns $H = 0$ as the time window is not full yet. At $t = t_A$, the first W samples have been acquired (the window is full) and a first estimation is provided, which is quite close to 25 s with this set of parameters. The estimation converges to the real inertia value at $t = t_B$, since the optimal value is used as the initial condition for MADS at the next time instant. We define the convergence time as $t_B - t_A$. Of course, if K is increased or different values of τ and Δ_0 are used, the convergence time can be 0, namely, the correct value of inertia is already obtained at t_A, which is a desirable feature.

In correspondence to the change of inertia, at $t = 250$ s, the estimation reaches the new value after a transient. Part of the transient (up to time t_C) is due to the inertia estimation algorithm itself, as all measurements inside the window must correspond to the new value of inertia, in order to obtain a correct estimation. The remaining delay (t_C-t_D) is instead due to the MADS algorithm and depends on parameters K, τ and Δ_0. This delay goes to zero when the convergence time is also zero. Both time intervals $[0, t_A]$ and $[250\text{ s}, t_C]$ have length $W\Delta t = 5$ s.

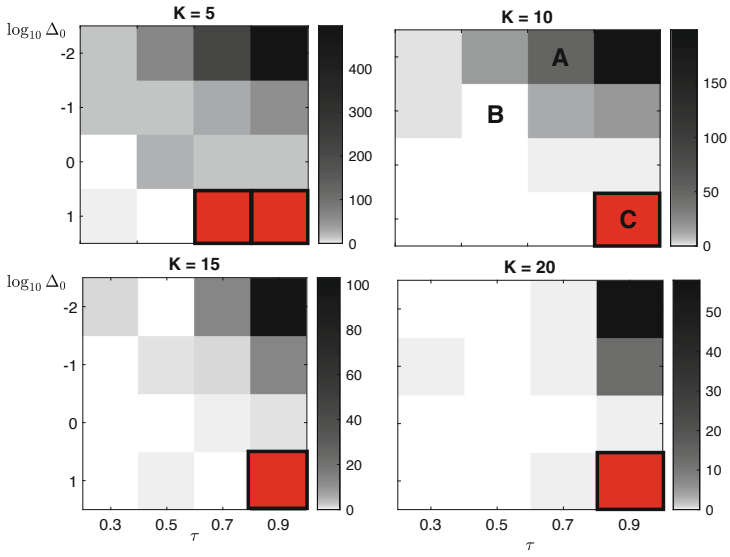


Fig. 5. Convergence time (from white to black) as a function of parameters τ, Δ_0, and K. The red rectangles correspond to parameter values leading to wrong estimations.

To assess the effect of parameters K, τ, and Δ_0 on the convergence time $t_B - t_A$, we performed several tests by setting $K = \{5, 10, 15, 20\}$, $\tau = \{0.3, 0.5, 0.7, 0.9\}$, and $\Delta_0 = \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$. Figure 5 shows the obtained convergence time (from white to black) for all considered combination of parameters. The white rectangles indicate the parameter settings for which the convergence time to a value close to the nominal one is exactly 0, whereas the red rectangles correspond to completely wrong estimations. We have made our analysis over parameters' ranges wide enough to include zero convergence time and completely wrong estimations, in order to explore accurately the parameter space and tune at best the algorithms. Parameter settings leading to both zero convergence time and a latency lower than 40 ms should be chosen. For example, Fig. 4 corresponds to the rectangle marked with A. Notice that, in general, low values of τ and high values of Δ_0 (bottom-left corner in the figure) ensure low convergence times. On the contrary, large values of both τ and Δ_0 (bottom-right corner) lead to a wrong behavior of the implemented algorithm. As soon as the number of iterations increases, several combinations of parameters ensure a null convergence time. By increasing K over a certain value, all rectangles in the figure would become white, testifying the robustness of the MADS algorithm.

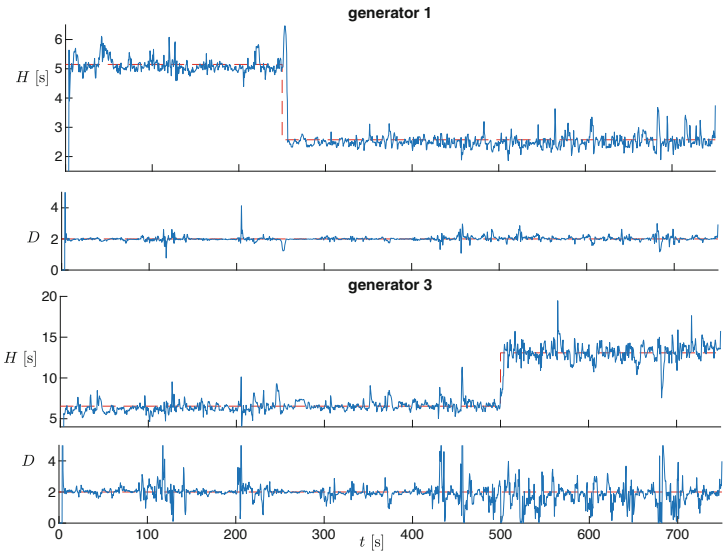


Fig. 6. Time evolution of the estimated inertia and damping of G1 and G3, obtained with $K = 10$, $\tau = 0.5$, and $\Delta_0 = 0.1$. The red dashed curves indicate the nominal values.

Figure 6 shows the estimation of H and D for both G1 and G3, with the set of parameters corresponding to rectangle B in Fig. 5. The microcontroller latency is 5.73 ms, which is much lower than both the sampling time Δt and

Table 2. Percent and RMS error on the steady-state inertia (for G1 and G3, before and after the step-change) and damping values for all generators.

Generator	Parameter	$\epsilon_{\%}$	ϵ_{rms}
G1	H (before)	–0.68	0.24 s
	H (after)	–2.42	0.22 s
	D	0.45	0.16
G2	H	0.19	0.53 s
	D	–3.36	0.40
G3	H (before)	–0.49	0.66 s
	H (after)	0.68	1.07 s
	D	–3.5	0.67
G6	H	7.45	0.87 s
	D	–4	0.34
G8	H	5.96	0.71 s
	D	–4.33	0.28 s

40 ms; the convergence time is 0 s. The oscillations are not due to the microcontroller implementation, but to the inertia estimation algorithm itself (see [10] for comparison), which relies on noisy measurements. They could be attenuated, e.g., by increasing the window length W, as deeply discussed in [10].

We computed the following errors on the steady-state values of the inertia (for G1 and G3, before and after the step-change) and of the damping, for all generators in the IEEE 14-bus power system:

$$\epsilon_{\%} = 100\frac{\bar{X} - X^{NOM}}{X^{NOM}}, \quad \epsilon_{rms} = \sqrt{\frac{1}{N_w}\sum_{i=1}^{N_w}(X_i - X^{NOM})^2} \tag{8}$$

with $\bar{X} = \frac{1}{N_w}\sum_{i=1}^{N_w} X_i$, where N_w is the number of time windows, X indicates either H or D and X^{NOM} the corresponding nominal values. The resulting errors are listed in Table 2 and are consistent with those reported in previous works employing similar algorithms [8,10,11]. The percent error is higher for G6 and G8 (even if we used $W = 20$ for these two generators), due to their much lower rated power, as also noticed in [10].

Finally, Fig. 7 corresponds to the rectangle C in Fig. 5. With this set of parameters, the implemented algorithm does not work properly, as the estimated inertia exhibits peaks up to the domain limit $H_{max} = 50$ s.

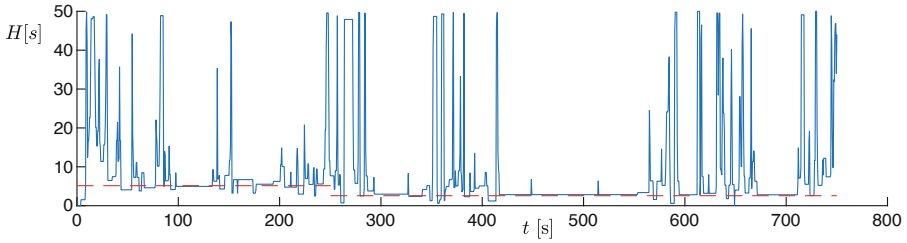


Fig. 7. Time evolution of the estimated inertia of G1, obtained with $K = 7$, $\tau = 0.9$, and $\Delta_0 = 10$. The red dashed curve corresponds to the nominal value.

4 Conclusions

We implemented an algorithm for the online estimation of inertia and damping in power grids on a low-cost microcontroller, by exploiting the MADS optimization algorithm. Considering that measurements of frequency are typically provided every 40 ms, a real-time estimation of H and D is possible, with delays of a few seconds, depending on the window length. With a sufficiently high number of iterations, the algorithm, converges to the correct inertia and damping values, independently of the choice of the other algorithm parameters. In future work, the embedded algorithm will be applied to the estimation of inertia and damping of a RES equipped with a synthetic inertia controller.

References

1. Kundur, P.: Power System Stability and Control. McGraw-Hill Inc. (1994)
2. Panda, R.K., Mohapatra, A., Srivastava, S.C.: Online estimation of system inertia in a power network utilizing synchrophasor measurements. IEEE Trans. Power Syst. **35**(4), 3122–3132 (2019)
3. Cai, G., Wang, B., Yang, D., Sun, Z., Wang, L.: Inertia estimation based on observed electromechanical oscillation response for power systems. IEEE Trans. Power Syst. **34**(6), 4291–4299 (2019)
4. del Giudice, D., Grillo, S.: Analysis of the sensitivity of extended Kalman filter-based inertia estimation method to the assumed time of disturbance. Energies **12**(3), 483 (2019)
5. Zhang, J., Xu, H.: Online identification of power system equivalent inertia constant. IEEE Trans. Industr. Electron. **64**(10), 8098–8107 (2017)
6. Cao, X., Stephen, B., Abdulhadi, I.F., Booth, C.D., Burt, G.M.: Switching Markov gaussian models for dynamic power system inertia estimation. IEEE Trans. Power Syst. **31**(5), 3394–3403 (2015)
7. Allella, F., Chiodo, E., Giannuzzi, G.M., Lauria, D., Mottola, F.: On-line estimation assessment of power systems inertia with high penetration of renewable generation. IEEE Access **8**, 62689–62697 (2020)
8. Zeng, F., Zhang, J., Chen, G., Wu, Z., Huang, S., Liang, Y.: Online estimation of power system inertia constant under normal operating conditions. IEEE Access **8**, 101426–101436 (2020)

9. Zhou, N., Pierre, J.W., Hauer, J.F.: Initial results in power system identification from injected probing signals using a subspace method. IEEE Trans. Power Syst. **21**(3), 1296–1302 (2006)
10. Baruzzi, V., Lodi, M., Oliveri, A., Storace, M.: Analysis and improvement of an algorithm for the online inertia estimation in power grids with RES. In: Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-5, May 2021
11. Makolo, P., Zamora, R., Lie, T.: Online inertia estimation for power systems with high penetration of RES using recursive parameters estimation. IET Renew. Power Gener. **15**, 2571–2585 (2021)
12. Anderson, P.M., Aziz A.: Power System Control and Stability. John Wiley & Sons, Fouad (2008)
13. Audet, C., Dennis Jr., J.E.: Mesh adaptive direct search algorithms for constrained optimization. SIAM J. Optim. **17**, 188–217 (2006)
14. Ravera, A., Oliveri, A., Lodi, M., Storace, M.: Embedded linear model predictive control through mesh adaptive direct search algorithm. In: Proceedings of the 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS), pp. 542-545, November 2019
15. Milano, F.: Power System Modelling and Scripting. Springer Science & Business Media (2010). https://doi.org/10.1007/978-3-642-13669-6
16. Bizzarri, F., Brambilla, A.: PAN and MPanSuite: simulation vehicles towards the analysis and design of heterogeneous mixed electrical systems. In: 2017 New Generation of CAS (NGCAS), pp. 1-4, September 2017

Random Weights Neural Network for Low-Cost Readout of Colorimetric Reactions: Accurate Detection of Antioxidant Levels

Edoardo Ragusa[1(✉)], Valentina Mastronardi[2], Deborah Pedone[2], Mauro Moglianetti[2,3], Pier Paolo Pompa[2], Rodolfo Zunino[1], and Paolo Gastaldo[1]

[1] Department of Electrical, Electronic, Telecommunication Engineering and Naval Architecture DITEN, University of Genoa, Genova, Italy
edoardo.ragusa@edu.unige.it

[2] Nanobiointeractions and Nanodiagnostics, Istituto Italiano di Tecnologia (IIT), Via Morego, 30–16163 Genova, Italy

[3] HiQ-Nano srl, Via Barsanti, 1-73010 Arnesano, Lecce, LE, Italy

Abstract. The introduction of Point Of Care (POC) devices is revolutionizing the field of diagnostics, thanks to their ease of use, portability, and real-time results. However, despite such advantages, POCs are still less accurate than traditional laboratory-based methods. In most cases, this is due to the qualitative on-off response of POCs along with readout procedures involving methods that are easily influenced by the environmental conditions or by the acquisition step of the result. Automation of the readout using machine learning supported by frugal devices and low-cost sensing systems can significantly enhance the quality of the analysis performed by POC devices, while maintaining the aforementioned advantages. This paper proposes the use of random-based neural networks to accurately assess the salivary antioxidant level detected through a colorimetric reaction. As a test case, a low-cost IoT device equipped with a trained neural network that infers the antioxidant level in the user's saliva was designed and tested. The experiments performed on real-world data confirm that the proposed solution outperforms the previously proposed readout strategy.

Keywords: Random weights neural networks · Point of care devices · Embedded systems

1 Introduction

Point Of Care (POC) diagnostics is a strongly growing technological field that has further gained interest in the latest two years of Covid-19 pandemics. POC diagnostics allows rapid on-site detection of infectious diseases as well as of several biomarkers of clinical relevance [1]. However, the integration of POC

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 95–104, 2023.
https://doi.org/10.1007/978-3-031-16281-7_10

systems with Internet of Things (IoT) and Machine Learning (ML) techniques could boost their diagnostic potential, providing quantitative data, more user-friendly and convenient operations, fully automated and intelligent detection, opening new perspectives for personalized medicine and preventive screenings [2].

In this framework, this paper reports a new solution for more effective and highly accurate POC colorimetric analysis. As a major contribution, this work provides a data-driven pipeline that is particularly suitable to reduce time to market and costs, making the proposed solution an appealing pipeline for commercial scenarios.

The proposal envisions random-based neural networks that feature a fast-training strategy, supported by a tight generalization bound that ensures good performance even when a few training samples are available, and leads to a prediction function that can be hosted by frugal electronics.

The proposed artificial intelligence solution has been integrated inside a low-cost IoT device that can quantify the amount of salivary antioxidants through the color change of the aqueous reaction hosted in a glass vial, using a low-cost RGB camera. This tool has the potential to significantly improve the performance of the POC sensor, without a significant hardware cost or increased complexity to the assay protocol.

This approach has been tested using, as a model case, the POC method proposed by Pedone et al. [3] for the rapid and visual evaluation of the body TAC through saliva, exploiting a colorimetric reaction between antioxidants and hydrogen peroxide catalyzed by Pt nanozymes. Figure 1 shows an example of the colorimetric outcome for two concentrations of antioxidants: the color intensity of the solution is indirectly proportional to the level of antioxidants, so the assay can be applied for the naked-eye measurement of the antioxidant level of the sample. The method proposed by Pedone et al. presented a semiquantitative readout strategy based on a smartphone camera. In this paper, we show a better accuracy compared with the original solution. In addition, we prove that random based neural networks (RBNs) compare favorably with other machine learning techniques for this analytical task.

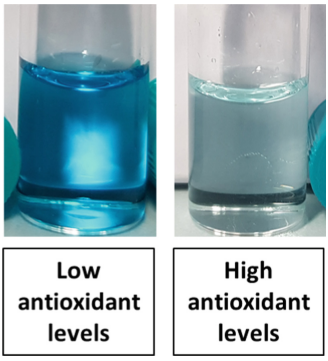


Fig. 1. POC device [3]: the color intensity denotes the level of antioxidants.

The paper is organized as follows. Section 2 briefly overviews the theoretical background of RBNs and reviews state-of-the-art works about the integration of POC devices with IoT and ML. Section 3 introduces the overall solution for accurate readout. Section 4 reports the outcomes of the experimental sessions. Finally, some concluding remarks are made in Sect. 5.

2 Preliminaries

In this section, we briefly present recent related works about the development of POC devices equipped with IoT/ML. Then we recall the basics of random-based neural networks.

2.1 Related Works

POC devices have recently attracted a surge in attention in various branches of research and different types have been developed for different applications [2,4,5] like lab-on-a-chip, lab-on-a-disc, microfluidic paper-based analytical [6], lateral flow devices, miniaturized PCR, and isothermal nucleic acid amplification [7].

Digital cameras have been playing a major role in low-cost sensing and to date various works have proposed such portable and easy-to-use low-cost solution as an alternative to spectrophotometric analyses [8,9].

However, the availability of low-cost sensing, processing, computing devices, and ML enables data transmission and processing data allowing accurate predictions and development of novel solutions [10]. Recently, POC devices have been proposed that allow you to remotely communicate with a centralized location for recording purposes that enable big data analysis [11,12]. Data login capability becomes fundamental considering that a large amount of data can be used to learn new rules from data automatically [13,14] and to aid the design effective sensing scheme [15].

Using machine learning for accurate readout is a well-known approach. Deep neural networks were employed in Ballard et al. [16] to jointly develop the CRP quantification algorithm with the multiplexed sensing membrane, selecting the most robust subset of sensing channels via feature selection methods. Lee et al. [17] proposed convolutional neural networks to infer spectrophotometric properties or the absorbance of the sample with limited sensing resources. Similarly, Luo et al. [18] presented a quantitative particle agglutination assay that combines mobile lens-free microscopy and deep learning for rapidly measuring the concentration of a target analyte. Yao et al. [9] presented a mobile spectrometer in combination with a Support Vector Regressor (SVR) designed to assess meat quality. Similarly, Chen et al. [19] employed a similar procedure for the aging analysis of rubber.

2.2 Random Based Neural Networks

The literature shows that RBNs provide a valuable tool for machine learning due to the simple training phase and good generalization performance. Many

variants of this approach have been presented in the last decades [20], namely Random Radial Basis Functions, Random Vector Functional-Link (RVFLs), Extreme Learning Machines (ELMs), and Weighted Sum of Random Kitchen Sinks.

The hypothesis space spanned by RBNs can be formalized as a weighted sum of non-linear functions:

$$y(\boldsymbol{x}) = \sum_{n=1}^{N} \beta_n \phi_n(\boldsymbol{x}) = \sum_{n=1}^{N} \beta_n \phi(\boldsymbol{w}_n \boldsymbol{x} + b_n) \tag{1}$$

where $\boldsymbol{x} \in R^D$ is the input pattern, $\boldsymbol{w}_n \in [-1; 1]^D$ and $b_n \in R$ are the parameters that characterize a neuron, $\boldsymbol{\beta} \in R^N$ is the vector of outputs' weights, and ϕ is a non-linear activation function $R^D \rightarrow R$.

In practice, RBNs use a simplified training procedure that sets a-priori random values for the parameters of the hidden neurons, $\boldsymbol{w}_n$ and b_n. Such aspect is pivotal when designing the learning strategy because the quality of the a-priori distribution sets the computing cost of the learning and prediction phases [21–23]. In addition, appropriate design strategies can lead to extremely convenient inference function that can be supported by ultra-constrained digital devices [24,25] making suitable almost any commercial microcontroller to host the inference phase.

Given a set of labeled data $\mathcal{T} = \{(\boldsymbol{x}_i, y_i), i = 1, ..., Z, \boldsymbol{x}_i \in [0, 1]^D, y_i \in [-1; 1]\}$, the training phase of RBNs consists in tuning the free parameters, β_n, to minimize a loss function. In general, the regularized mean square error is adopted:

$$\min_{\boldsymbol{\beta}} \{\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2\} \tag{2}$$

When $Z \leq N$, one has:

$$\boldsymbol{\beta} = \boldsymbol{H}^T (\lambda \boldsymbol{I} + \boldsymbol{H}\boldsymbol{H}^T)^{-1} \boldsymbol{y} \tag{3}$$

conversely, when $Z > N$ one has:

$$\boldsymbol{\beta} = (\lambda \boldsymbol{I} + \boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{y} \tag{4}$$

where $\boldsymbol{H}$ denote a $Z \times N$ matrix with $h_{in} = \phi_n(\boldsymbol{x}_i, \boldsymbol{w}_n, b_n)$. The quantity λ is a regularization term that rules the trade-off between data fitting and the smoothness of the solution. It is worth noting that (3) and (4) can be solved by exploiting computationally efficient algorithms for the solution of linear equation systems, thus avoiding explicit matrix inversions.

3 Case Study: Accurate Measure of Antioxidant Level in Saliva Using Colorimetric Reactions

To measure the antioxidant level in saliva a drop of saliva was added to a predosed aqueous solution of platinum nanoparticles, followed by the chromogenic

solution and hydrogen peroxide. After 5 min at room temperature, the color change was quantified by measuring the absorbance at 652 nm by NanodropC spectrophotometer and by our proposed POC automated device. Despite the quick and easy method, we can identify two major aspects in the design of an accurate readout strategy for colorimetric POC devices hosted in glass test tubes. First, the picture should be standardized. Second, the link between the color and the quantity of interest should be accurately retrieved.

Smartphone-based measurements of colorimetric response are commercially appealing because smartphones are widespread. However, this kind of solution requires the user to take care of the overall setup: lighting conditions like intensity, illuminations, shadows, and direction. In addition, pictures are taken from various distances and framing angles.

A support that equalizes all the measures, acting like a spectrophotometer in the visible range, can significantly reduce dispersion, enhancing overall accuracy and usability. To this aim, we designed a simple 3D printed support (Fig. 2) of size 10 × 6 × 6 cm, equipped with low-cost electronics. The device is "L" shaped: in the first branch a white led sets the lighting level with frequency (6000–6500K), the test tube is placed into the center of the corner of the support, and a camera is placed at right angle in the other branch to record the color changes.

(a) (b)

Fig. 2. Experimental setup (a) Schematic representation of the device. (b) Low cost readout system equipped with frugal electronics: in the top-left corner is placed the led. The camera and the embedded system are hosted in the bottom right corner.

In this setup, the camera is interfaced with an ESP32 module. This device equips a high-end micro-controller and a Wifi/Bluetooth module to stream the captured images to a web server. Eventually, a desktop computer performs the inference procedure. It should be noted that the system has been designed to record training data. In a commercial version, the ESP32 module could be replaced by a middle-end micro-controller that captures the image and performs the inference operation directly on site.

Figure 3 shows an example of a low-quality acquisition by a cheap digital camera. The part of interest marked by the red rectangle is isolated from the

background and the average color value is computed for each channel. A patch from the background (the green rectangle) is isolated. The average value is computed for the pixels of this patch too. Eventually, the system processes the difference between the average value of the patch of interest and the reference tag. In practice, the images are represented as vectors in R^3, where each component of the vector identifies one of the three RGB components.

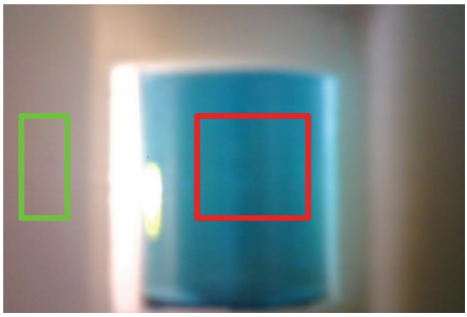


Fig. 3. Example of a low-quality frame captured by the readout device. The red rectangle limits the pixels that convey the colorimetric information analyzed. The green rectangle identifies the reference patch.

A previously trained RBN network receives the measured color and outputs the estimated level of antioxidants in the user's saliva in the form of a scalar value.

4 Experiments

The experimental campaign assesses the performance of the proposed system for accurate readout. Firstly, the dataset is described. Later the accuracy of the proposed approach is compared against alternative solutions.

A dataset composed of 464 samples has been collected exploiting data from 8 users. We collected 58 measures per user, one each five-second. A measure involves a triple RGB and the corresponding label. Every measure has been compared with the absorbance value recorded using a laboratory spectrophotometer. The measures were taken in parallel using the two devices on the same pool of material. One could argue that, in ideal settings, a known relationship links RGB values and absorbance. However, to use the explicit formula, one should take into account multiple factors, for example, the exact light frequency of the led is unknown due to parameters dispersion of the real devices, camera parameters and distorsion introduced by digital representation should be explicitly modeled. Interested reader can refer to [26] for the details.

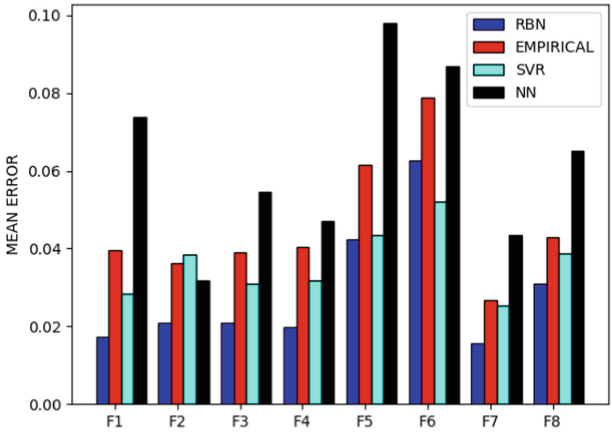
On the one hand, random-based networks are well-known to be efficient but, on the other hand, concerns about generalization performance raises in a few

applications. To evaluate the generalization effectiveness of RBNs to support accurate colorimetric quantization we compared RBNs with two regression methods, namely a Support Vector Regressor with gaussian kernel (SVR) and a fully connected neural network (NN). We opted for these two baseline solutions due to their promising use in recent works for similar tasks [9,16,18]. In addition, we compared the results with the empirical model (EMPIRICAL) presented in [3] for smartphone inference.

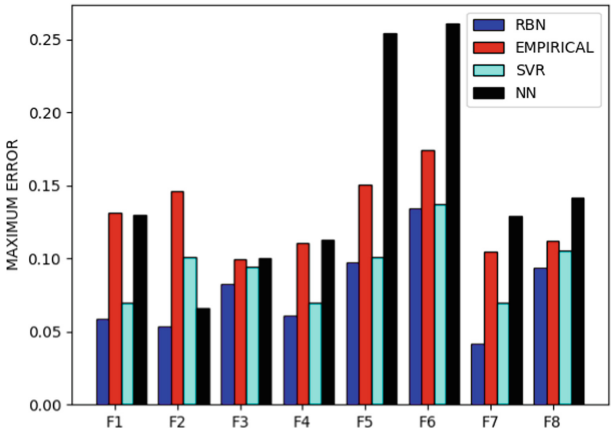
We employed a standard 8-fold validation procedure where every fold contains data of one user only. Accordingly, 8 training procedures were performed using 7 folds as training set and 1 fold as the test set. Hyper-parameters were set using a standard grid search using 33% of the training data as validation set, therefore testing data were never involved in the validation procedure. Following standard practice, we averaged the results for RBNs over 20 random extractions of the hidden layer. For the RBN the hyper-parameters selection involved the number of hidden neurons and regularization parameter λ. Similarly for SVR were selected $\{\sigma, C\}$, for NN the number of hidden neurons, the number of hidden layers, and the learning rate. Eventually, the NN was trained using Stochastic gradient descent with a maximum number of epochs equal to 1000 and early stop regularization with patience set to 5.

Figure 4 summarizes the performance in terms of the mean absolute error of the proposed solution against the baselines. The figure is divided into bar plots, one for each fold from F1 to F8. The bar plots contain four-bar that report the mean absolute error computed for the test folder for every method tested.

The results highlight that, for the test under analysis, RBNs outperform other solutions in most of the cases. Only in folder F6, the SVR outperforms the RBNs. In general, the distance between the inferred absorbance and the actual value is small. In 5 folds over 8, the average error is lower than 0.02., i.e. lower than 2% of the admissible range of absorbance. In fold 6 one can observe that all the methods scored a larger average error. Probably, this sample generated an anomalous distribution that affected all the 58 testing patterns of fold F6. One could note that SVR and NN seem to have sub-optimal performance. However, there are a few justifications. First, the input dimensionality is extremely small, i.e. input patterns belong to R^3. Second, the dataset is characterized by a modest amount of training data 464-58 because every time a fold is removed. Third, it is likely that the contribution of the three channels is not equal. In fact, the colorimetric reaction mostly affects the blue spectra, indeed, green and red channels are expected to convey a modest amount of information. This becomes important if one notes that SVR uses Euclidean distance that equalizes the contribution of the three channels. Finally, the NN is affected by a non-convex optimization problem. This leads to a more difficult optimization problem that in some cases fails in converge to good solutions. In the experiment, the methods shared fully data-driven optimization where hyper-parameters were set using a greed search. It is feasible to think that a dedicated training procedure could boost the NN performance.



(a) Results: average error on the testing fold.



(b) Results: maximum error on the testing fold.

Fig. 4. Experimental results: analysis of the generalization performance.

Figure 4 inherits the format of Fig. 4 but on the y-axis, it is shown the maximum error computed on the test fold, i.e. the maximum discrepancy between the nanodrop measure and the inferred value from RGB. The trend is very close to the ones shown for average values.

5 Conclusion

The present paper described an application of pervasive intelligence where a machine learning enables low cost and ability of on-site analysis, but, often prove less accurate when compared to traditional laboratory tests and assays. In most cases this is due to readout procedure that involves non-skilled users. Automation of the readout procedure using machine learning supported by frugal devices and low-cost sensing systems can significantly enhance the quality of analysis performed by POC devices maintaining the aforementioned advantages. This paper proposes the use of random-based neural networks to accurately assess the antioxidant level detected by a recently proposed colorimetric reaction. As a test case, a low-cost IoT device equipped with a trained neural network that infers the antioxidant level in user saliva was designed and tested. The experiments performed on real-world data confirm that the proposed solution outperforms the previously proposed readout strategy.

References

1. Heidt, B., et al.: Point of care diagnostics in resource-limited settings: a review of the present and future of POCc in its most needed environment. Biosensors **10**(10), 133 (2020)
2. Wang, C., Liu, M., Wang, Z., Li, S., Deng, Y., He, N.: Point-of-care diagnostics for infectious diseases: from methods to devices. Nano Today **37**, 101092 (2021)
3. Pedone, D., Moglianetti, M., Lettieri, M., Marrazza, G., Pompa, P.P.: Platinum nanozyme-enabled colorimetric determination of total antioxidant level in saliva. Analyt. Chem. **92**(13), 8660–8664 (2020)
4. Papadakis, G., et al.: Portable real-time colorimetric lamp-device for rapid quantitative detection of nucleic acids in crude samples. Sci. Rep. **12**(1), 1–15 (2022)
5. Mastronardi, V., Moglianetti, M., Ragusa, E., Zunino, R., Pompa, P.P.: From a chemotherapeutic drug to a high-performance nanocatalyst: a fast colorimetric test for cisplatin detection at ppb level. Biosensors **12**(6), 375 (2022)
6. Pomili, T., Donati, P., Pompa, P.P.: Based multiplexed colorimetric device for the simultaneous detection of salivary biomarkers. Biosensors **11**(11), 443 (2021)
7. Tatulli, G., Pompa, P.P.: An amplification-free colorimetric test for sensitive DNA detection based on the capturing of gold nanoparticle clusters. Nanoscale **12**(29), 15604–15610 (2020)
8. Cai, F., Lu, W., Shi, W., He, S.: A mobile device-based imaging spectrometer for environmental monitoring by attaching a lightweight small module to a commercial digital camera. Sci. Rep. **7**(1), 1–9 (2017)
9. Yao, X., Cai, F., Zhu, P., Fang, H., Li, J., He, S.: Non-invasive and rapid PH monitoring for meat quality assessment using a low-cost portable hyperspectral scanner. Meat Sci. **152**, 73–80 (2019)
10. Mayer, M., Baeumner, A.J.: A megatrend challenging analytical chemistry: biosensor and chemosensor concepts ready for the internet of things. Chem. Rev. **119**(13), 7996–8027 (2019)
11. Yang, T., Gentile, M., Shen, C.-F., Cheng, C.-M.: Combining point-of-care diagnostics and internet of medical things (IOMT) to combat the covid-19 pandemic (2020)

12. Alonso, O., et al.: An internet of things-based intensity and time-resolved fluorescence reader for point-of-care testing. Biosens. Bioelectron. **154**, 112074 (2020)
13. John-Herpin, A., Kavungal, D., von Mücke, L., Altug, H.: Infrared metasurface augmented by deep learning for monitoring dynamics between all major classes of biomolecules. Adv. Mater. **33**(14), 2006054 (2021)
14. Gadalla, A.A., Friberg, I.M., Kift-Morgan, A., Zhang, J., Eberl, M., Topley, N., Weeks, I., Cuff, S., Wootton, M., Gal, M., et al.: Identification of clinical and urine biomarkers for uncomplicated urinary tract infection using machine learning algorithms. Sci. Rep. **9**(1), 1–11 (2019)
15. Ballard, Z.S., Shir, D., Bhardwaj, A., Bazargan, S., Sathianathan, S., Ozcan, A.: Computational sensing using low-cost and mobile plasmonic readers designed by machine learning. ACS Nano **11**(2), 2266–2274 (2017)
16. Ballard, Z.S., et al.: Deep learning-enabled point-of-care sensing using multiplexed paper-based sensors. NPJ Digit. Med. **3**(1), 1–8 (2020)
17. Lee, J., et al.: Machine learning promoting extreme simplification of spectroscopy equipment, arXiv preprint arXiv:1808.03679 (2019)
18. Luo, Y., Joung, H.-A., Esparza, S., Rao, J., Garner, O., Ozcan, A.: Quantitative particle agglutination assay for point-of-care testing using mobile holographic imaging and deep learning. Lab Chip **21**(18), 3550–3558 (2021)
19. Chen L., Chen, X., Li, X., Fu, X., Zhang, R., Wang, T.: Determine the aging status of silicone rubber insulators base on smartphone hyperspectral cameras. In: 2021 International Conference on Power System Technology (POWERCON), pp. 2399–2403. IEEE (2021)
20. Cao, W., Wang, X., Ming, Z., Gao, J.: A review on neural networks with random weights. Neurocomputing **275**, 278–287 (2018)
21. Dudek, G.: A constructive approach to data-driven randomized learning for feedforward neural networks. Appl. Soft Comput. **112**, 107797 (2021)
22. Dudek, G.: Data-driven randomized learning of feedforward neural networks. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
23. Ragusa, E., Gastaldo, P., Zunino, R., Cambria, E.: Balancing computational complexity and generalization ability: a novel design for ELM. Neurocomputing **401**, 405–417 (2020)
24. Ragusa, E., Gianoglio, C., Zunino, R., Gastaldo, P.: A design strategy for the efficient implementation of random basis neural networks on resource-constrained devices. Neural Process. Lett. **51**(2), 1611–1629 (2020)
25. Ragusa, E., Gianoglio, C., Zunino, R., Gastaldo, P.: Random-based networks with dropout for embedded systems. Neural Comput. Appl. **33**(12), 6511–6526 (2020). https://doi.org/10.1007/s00521-020-05414-4
26. Soda, Y., Robinson, K.J., Cherubini, T.J., Bakker, E.: Colorimetric absorbance mapping and quantitation on paper-based analytical devices. Lab Chip **20**(8), 1441–1448 (2020)

Resource-Constrained Implementation of Deep Learning Algorithms for Dynamic Touch Modality Classification

Haydar Al Haj Ali[1(✉)], Christian Gianoglio[1], Ali Ibrahim[1,2], and Maurizio Valle[1]

[1] Department of Electrical, Electronic, Telecommunication Engineering and Naval Architecture, DITEN University of Genoa, Genova, Italy
{Haydar.AlHajAli,christian.gianoglio}@edu.unige.it, maurizio.valle@unige.it

[2] Department of Electrical and Electronics Engineering, Lebanese International University (LIU), Beirut 1105, Lebanon
ali.ibrahim@liu.edu.lb

Abstract. Integrating Machine Learning (ML) algorithms with tactile sensing arrays yield sophisticated systems capable of performing intelligent tasks. Such systems can be used in prosthetic devices and robotics applications, enabling conducting daily tasks and manipulations. This paper presents low-cost and resource-constrained implementations of deep learning algorithms for the classification of dynamic touch modality based on alphabetic letter patterns. This work provides a comparison between two types of deep neural networks: 1-D Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Moreover, the models providing the best performance in terms of accuracy and computational cost have been deployed on a resource-constrained embedded system. Experimental results show that 1-D CNN outperforms RNNs in terms of both accuracy and computational cost achieving a classification time of 242 ms using 32-bit floating point on the Arduino Nano BLE hardware device.

Keywords: Tactile sensing system · Recurrent neural networks · Convolutional neural networks · Embedded neural networks · Computational cost

1 Introduction

Tactile sensing systems represent an attractive field of study in the last years due to their capabilities in emulating the human skin. This impressive trait of the tactile sensing system can be exploited in many application domains such as robotics and prosthetic devices [1–3]. Generally, a tactile sensing system is comprised of three main parts: 1) an array of tactile sensors to capture and convert the mechanical stimulus, 2) an interface electronics for signal conditioning and data acquisition, and 3) an embedded digital processing unit for tactile data decoding.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 105–115, 2023.
https://doi.org/10.1007/978-3-031-16281-7_11

Incorporating Artificial Intelligence (AI) models with tactile sensing system yields in enabling intelligent tasks such as: texture discrimination, touch modality classification, slippage detection and, object recognition [4–6]. To elaborate more, front-end tactile sensors have to acquire particular parameters from their surfaces such as force, vibration, and impedance. Such information will be analyzed by either machine learning (ML) models in an indirect manner by means of extracting hand-crafted features or by using more powerful tools based on deep learning networks that extract directly features from data. However, despite the advances in this field, embracing tactile sensing systems in real applications is challenging and encounters many limitations especially when it comes to the network deployment part [7]. These limitations are related to the complexity of the processing algorithms and the massive number of parameters in the deep networks, that will affect the energy consumption and latency of the hardware devices.

From this perspective, this work presents a dynamic touch modality based on alphabetic letters patterns classification framework using recurrent neural networks and 1-D Convolutional neural networks. The goal of this work is to find a trade-off implementation among different sound algorithms to achieve an efficient alphabetic letter patterns classification model to be deployed on the edge. Real-time tactile data are characterized by its 3-dimensional tensor structure, in which the first two dimensions represent the geometry of the sensors array, whereas the third dimension is affiliated to the time. Therefore, the aforementioned networks were targeted in this work due to their capabilities in processing time-series data. The main contributions of this work are as follows:

- Explore the performance of 1-D CNN and two RNNs (long short term memory (LSTM) and Gated Recurrent Unit (GRU)) in the classification of dynamic touch modalities based on alphabetic letter patterns.
- Conduct a comparison between the aforementioned models and select an efficient network in terms of accuracy and hardware complexity i.e. number of parameters and Floating Point Operations (FLOPs).
- Deploy the chosen network on a low-cost Arduino Nano 33 BLE microcontroller and achieve real-time functionality.

2 Related Works

One of the crucial motivations for studying the human sense of touch is its vital role in our interactions and the manipulation of our environment. The main goal is to convey this human feature to robotics and prosthetic devices to enable more sophisticated tasks and make such systems plausible to perform daily tasks.

Christian et al. [4] employed 1-D CNN to discriminate between three touch modalities (slide a finger, brush a paintbrush, and roll a washer) acquired from a 4×4 piezoelectric tactile sensors. They conducted down-sampling on the raw sensor signals using four different time windows to reduce the dimension of the input tensor. The authors also deployed the best-performing model on Arduino Nano 33 BLE micro-controller, achieving an inference time of 128 ms using 32-bit

floating-point and 26 ms using the quantized 8-bit integer representations. In [8], the same three touch modalities were classified using Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). The two models achieved remarkable accuracies of 84.23% and 83.78% for LSTM and GRU respectively. Those results yielded also a reduction of 99.989% in the number of FLOPs compared to the same problem in the literature [9]. In [10] nine static touch modalities (tap, pat, push, stroke, scratch, slap, pull, squeeze, and no-touch) collected from 40 participants, are classified using LogitBoost. The model achieved a classification accuracy of 71%.

Alameh et al. [11] transformed the 3D tensorial tactile data to 2D RGB images and used pre-trained convolutional neural networks (CNN) to perform classification. The authors adopted an Inception Resnet model resulted in a notable classification accuracy (76.9%) but implied an excessive computational cost. In [12] the authors employed different algorithms to classify 14 touch gestures. Results achieved reported that kernel-SVM (Support Vector Machine) attained the highest accuracy among others. In [13], the authors adopted the AlexNet to classify 22 objects from pressure images acquired from tactile sensor array. The authors deployed the model on three different hardware devices: Movidius NCS2, Raspberry Pi4 and Jetson TX2. The results reported that Jetson TX2 achieved the smallest power consumption and inference time. In [14] two different approaches were adopted to classify eight objects (finger, hand, arm, pen, scissors, pliers, sticky tape and Allen-key) using 28×50 tactile sensory array. The first approach depends on Speeded-Up Robust Features (SURF) descriptor, while the second approach uses a pre-trained AlexNet CNN. Features acquired from both approaches were used to train Support Vector Machine (SVM). Ibrahim et al. [6] involved deep neural networks trained with transfer learning to classify 22 objects using pressure images acquired from 16×10 piezoresistive sensors. Adopting the AlexNet model yielded a remarkable classification accuracy of 99.5%. In [15], the authors adopted single-layer feed-forward neural networks (SLFNN) based on Extreme Learning Machine (ELM) to classify the three touch modalities presented in [4]. They defined three different feature sets and compared the performance of the networks on each in terms of accuracy and FLOPs cost. The work in [16] targeted the classification of three objects with different stiffness (Low, Medium, High) acquired from 4×2 piezoelectric sensors. The authors employed a single-layer feed-forward network (SLFNN) with five different sizes of hidden layer (25, 50, 100, 200, 300) using statistical features (Standard Deviation, Mean, Skewness, Kurtosis). Yielded results showed a classification accuracy of 95.5% when 50 hidden neurons were adopted in the network.

3 Deep Networks for Touch Modality Classification

In general, real-time data are considered Spatio-temporal data, as the temporal dimension represents the duration of the applied variable, while the spatial aspect defines the space where this action took place. In our case, this is applicable on our target that is represented by dynamic touch modalities based on alphabetic

letter patterns. As mentioned in the introduction, deep networks are capable of extracting features in a direct manner from time-series data, without the need of crafting the features by hand. Accordingly, three different deep neural networks have been selected and implemented to classify the touch modalities namely 1-D CNNs, LSTM, and GRU.

3.1 1-D Convolutional Neural Networks

Employing CNNs for tactile application [8,11] imposes a pre-conversion process of time-series data into image representations, to be fed to the network. However, 1-D CNN is capable of processing 1-D time sequences. Such classifiers can process 3D input tensor while performing convolution operation in one dimension [4]. The convolution operation is conducted by 3D kernels that pass on the input tensor and slide along the time direction yielding a vector of features. Thus, the network learns the Spatio-temporal representation of the input tensor.

3.2 Recurrent Neural Networks

RNNs are capable of modeling long-range temporal dependencies [8]. Dynamic touch modalities are wealthy with temporal features, that make each class characterized by its particular temporal behavior, thus making RNNs a potential candidate for this classification task. In this work, we chose two recurrent networks based: LSTM and GRU both having similar working methodology, enabling them to model long-range temporal dependencies [17]. The main distinction between aforementioned networks is the degree of complexity, as the GRU is less complex than LSTM since it involves fewer number of gates in its architecture.

4 Experimental Setup

4.1 Dataset

The dataset is composed of twenty-six alphabetic letter patterns from A to Z. Five subjects have participated in this experiment, in which each participant has been asked to perform ten trials for each class on the top surface of the tactile sensor array described in the following. For each trial, the acquisition time was set to 4 s at a sampling rate of 30 samples/second. The resulting dataset is composed of 1300 3D-tensor data where the first two dimensions are the geometry of the tactile array while the third dimension represents the time slot (4 s) as shown in Fig. 1.

The tactile sensing system involves three main components which are: 1) a tactile sensor array, 2) an electronic interface, and 3) a graphical user interface on a computer. Figure 2 shows the block diagram of the system used in this work, along with an illustration of the collected dataset. The different blocks of the system are explained as follows:

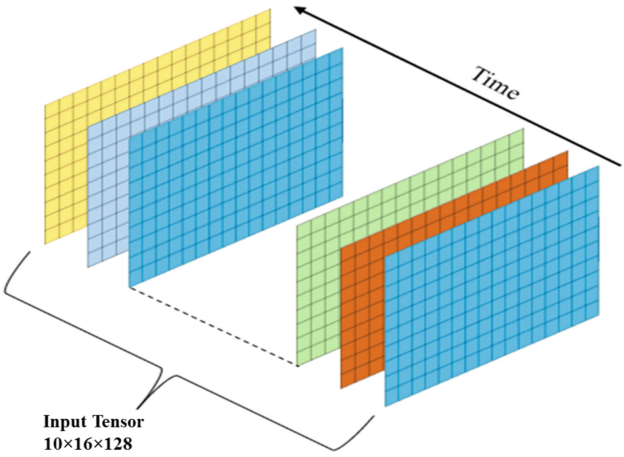


Fig. 1. Input 3D-tensor (1 Sample).

1. **Tactile sensor array**: the MS9723 FSR piezoresistive sensor [18] is adopted as a sensor array in this work. The MS9723 is composed of 10 rows $\times$ 16 columns forming a total of 160 sensors converting the applied pressure into electric resistance variation. This variation of resistance yields current with a proportional value of the applied force.
2. **Snowboard system**: it is employed as an electronic interface for the system. The snowboard system is developed by Kitronyx and it is considered as a resistive version of modern capacitive multi-touch sensing technology [19]. This system conveys the detected forces by the MS9723 sensors to the computer through USB interface.
3. **Graphical User Interface**: the GUI is based on Snowforce 3, which is a tool to communicate between the snowboard and the computer [20]. This GUI manages the communication parameters, the visualization, and the data logging period.

4.2 Implementation

The training strategy adopted in this work is based on model selection, in an attempt to deploy the best model among 1-D CNN and RNNs, in terms of achieved accuracy and computational cost computed as the number of parameters and FLOPs. Different architectures were trained per model, depending on hyperparameter tuning for the classifiers architecture. More discussion is found in the subsections below.

1-D CNN Classifier: Different 1-D CNN architectures have been examined by changing the number of convolutional layers, filters, and kernel size. The motivation behind this is to examine the impact of the aforementioned parameters on the accuracy and the computational cost as well.

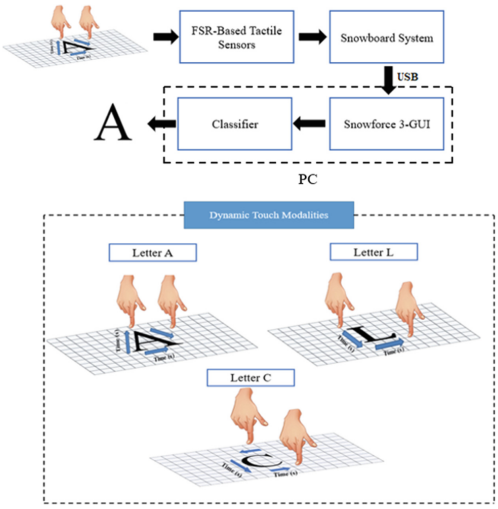


Fig. 2. Illustration of the tactile sensing system and the dynamic touch modality classification scenario.

1. Number of convolutional layer from 2 to 3.
2. Filter candidates are: (4, 4), (4, 8), (8, 8), (4, 4, 4), (8, 8, 8), (4, 8, 16).
3. Kernel size: $\{4, 8, 10\}$.

To explain more, f = (4,8,16) identifies the implementation of 3 convolutional layers along with the application of 4, 8, and 16 kernels to the first, second, and third convolutional layer, respectively. Thus, the experiment includes the implementation of 18 CNN models covering all the possible formulations.

LSTM and GRU Networks: The two networks are composed of a single hidden layer but with a set of 5 different sizes: $\{10, 20, 30, 40, 50\}$ (number of neurons in the hidden layer). Therefore, 5 different architectures per model have been explored in this work.

All the models were implemented and trained in Python by using Keras API in Tensorflow, with the following settings: number of epochs = 200, patience = 10 (early stop criterion), batch size = 20, and learning rate of 0.01. A stratified 5-folds were generated, in which 4 folds were used in turn for training and validation (for model selection) by a ratio of 80/20% per fold, and 1 fold among the rest is used for testing (to estimate generalization performance). The Adam optimizer was adopted to train all the networks, along with categorical cross-entropy as a loss function and softmax activation for LSTM and GRU output layer and ReLu activation for CNN. Prior starting the training, the input tensors were rearranged from $10 \times 16 \times 128$ to 160×128 by the input flatten layer for all the networks.

Embedded Hardware Device: The model to be deployed is chosen on the best trade-off between generalization accuracy and computational cost. The deployment is done on an Arduino Nano 33 BLE, as it could be embedded in prosthetic devices and robotic systems. The Arduino Nano 33 BLE is a low-cost, resource-constrained device having a CPU clock frequency of 64 MHz, a flash memory size of 1 MB, SRAM holding 256KB, and an operation voltage of 3.3 V.

5 Experimental Results

Table 1 shows the best accuracies achieved by the three models (1-D CNN, GRU and LSTM) among the 5 folds, along with the associated computational parameters for each. Obtained results report that 1-D CNN outperforms RNNs in terms of achieved accuracy on the test sets (97.85% for 1-D CNN, 87.30% for LSTM and 85.77% for GRU) and from the computational cost side measured as number of parameters and FLOPs. To add more, the table illustrates the accompanied architecture that yields the best accuracy over the folds. For instance, the architecture with 3 convolutional layers of filter size 8 and kernel 4 ($f = [8, 8, 8], k = 4$) scored the highest accuracy on $Fold0$ (97.31%) among the rest of the 18 1-D CNN models described above, with a total of 5,890 parameters and 137.1 kFLOPs involved in the model. The same discussion is applicable among all the five folds for the CNN.

As for LSTM, the architecture with a hidden layer of size 50 possessed the best accuracy among all the folds except $Fold3$ (achieved by a hidden layer of size 40). GRU network (hidden layer size 40) scored the highest accuracy among $Folds$ 0, 1 and 4, while $Folds$ 2 and 3 were affiliated to a hidden layer of size 50 and 30 respectively. The implementation on the hardware device demands a trade-off between accuracy and complexity. In this study, CNNs outperformed LSTMs and GRUs in terms of both accuracy and computational cost. In particular, CNNs achieved higher accuracy (∼10%) with lower FLOPs (∼11 k) and model parameters. Thus, we can conclude that 1-D CNN represents the optimal candidate to be deployed on the target hardware device. Table 3 shows the average accuracy, models parameters and FLOPs for all the 18 CNN architectures explored in this work. The implementation on the Arduino Nano 33 BLE involved 3 chosen models: the model with lowest FLOPs and Parameters (CNN0 Parms = 2,762, FLOPs = 673.8 k), the model with highest average accuracy (CNN7 avg acc = 97.70%), and the model aggregating a compromise between accuracy and complexity (CNN15 avg acc = 94.30%, Parms = 3,670, FLOPs = 721.8 k). The three models were exported in tflite using the float32 representation.

Table 1. Comparison of accuracy and number of parameters

	1-D CNN			LSTM			GRU		
Folds	Avg Acc	Model Prm	kFLOPs	Avg Acc	Model Prm	kFLOPs	Avg Acc	Model Prm	kFLOPs
Fold0	.973 $[8, 8, 8]$ $k = 4$	5,890	137.1	.865 $n = 50$	43,526	11291.9	.862 $n = 40$	25,306	6750.0
Fold1	.973 $[8, 8, 8]$ $k = 8$	11,522	273.1	.885 $n = 50$	43,526	11291.9	.869 $n = 40$	25,306	6750.0
Fold2	.981 $[8, 8]$ $k = 8$	5,626	270.4	.862 $n = 50$	43,526	11291.9	.854 $n = 50$	25,306	8885.6
Fold3	.9769 $[4, 8]$ $k = 8$	5,622	136.1	.873 $n = 40$	33,226	8623.9	.831 $n = 30$	25,306	4793.7
Fold4	.989 $[8, 8, 8]$ $k = 10$	14,338	3410.5	.891 $n = 50$	43,526	11291.9	.873 $n = 40$	25,306	6750.0
Avg Acc $\pm$ **Std**	$.979 \pm .005$			$.873 \pm .013$			$.858 \pm .015$		

Table 2. Inference time (ms)

Model	Inference time (ms)
CNN0	242
CNN7	895
CNN15	260

Table 2 provides the average inference time on 100 data. As expected, the model with the highest number of parameters (CNN7) scores the highest inference time 895 ms, while the rest of the models achieved convergent inference time due to their low computational cost. Therefore, the timing performance in CNN0 and CNN15 met the real-time constraints (400 ms) of perceiving a touch based on a psychometric test done in [21].

Table 3. Average accuracy and parameters for all CNN architectures

Model	Average accuracy ± Stdev	Model parameters	kFLOPs
CNN0 [4,4] k = 4	**.847 ± .052**	**2,762**	**673.8**
CNN1 [4,4] k = 8	.837 ± .100	5,386	1337.3
CNN2 [4,4] k = 10	.840 ± .068	6,698	1669.1
CNN3 [4,8] k = 4	.938 ± .017	2,934	688.7
CNN4 [4,8] k = 8	.952 ± .023	5,622	1360.5
CNN5 [4,8] k = 10	.941 ± .024	6,966	1696.4
CNN6 [8,8] k = 4	.963 ± .017	5,626	1360.7
CNN7 [8,8] k = 8	**.977 ± .008**	11,002	**2704.2**
CNN8 [8,8] k = 10	.962 ± .017	13,690	3376.0
CNN9 [4,4,4] k = 4	.760 ± .080	2,830	674.6
CNN10 [4,4,4] k = 8	.821 ± .086	5,518	1342.3
CNN11 [4,4,4] k = 10	.819 ± .078	6,862	1676.1
CNN12 [8,8,8] k = 4	.945 ± .018	5,890	1370.6
CNN13 [8,8,8] k = 8	.955 ± .018	11,522	2730.5
CNN14 [8,8,8] k = 10	.970 ± .021	14,338	3410.4
CNN15 [4,8,16] k = 4	**.943 ± .041**	**3,670**	**721.8**
CNN16 [4,8,16] k = 4	.96 ± .023	6,870	1426.3
CNN17 [4,8,16] k = 4	.950 ± .013	8,470	1778.5

6 Conclusion

In this paper, we explored the potential of two types of deep neural networks: Recurrent and Convolutional Neural Networks, for dynamic touch modality classification based on alphabetic letter patterns. The results report that 1-D CNN outperforms Recurrent neural networks (LSTM and GRU) in terms of achieved accuracy and computational cost. Eventually, we deployed the 1-D CNN model on a low-cost resource-constrained hardware device (Arduino Nano 33 BLE) by using a 32-bit floating-point and achieving real time functionality. In conclusion, the proposed work provided an efficient embedded deep neural network implementation for a tactile sensing system that can be adopted in robotics and prosthetic device applications.

References

1. Dahiya, R.S., Mittendorfer, P., Valle, M., Cheng, G., Lumelsky, V.J.: Directions toward effective utilization of tactile skin: a review. IEEE Sens. J. **13**(11), 4121–4138 (2013)
2. Cheng, G., Dean-Leon, E., Bergner, F., Olvera, J.R.G., Leboutet, Q., Mittendorfer, P.: A comprehensive realization of robot skin: sensors, sensing, control, and applications. Proc. IEEE **107**(10), 2034–2051 (2019)
3. Abbass, Y., Saleh, M., Dosen, S., Valle, M.: Embedded electrotactile feedback system for hand prostheses using matrix electrode and electronic skin. IEEE Trans. Biomed. Circuits Syst. **15**(5), 912–925 (2021)
4. Gianoglio, C., Ragusa, E., Zunino, R., Valle, M.: 1-D convolutional neural networks for touch modalities classification. In: 2021 28th IEEE International Conference on Electronics, Circuits, and Systems (ICECS), pp. 1–6. IEEE (2021)
5. Su, Z., et al.: Force estimation and slip detection/classification for grip control using a biomimetic tactile sensor. In: 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), pp. 297–303 (2015)
6. Ibrahim, A., Ali, H.H., Hassan, M.H., Valle, M.: Convolutional neural networks based tactile object recognition for tactile sensing system. In: Saponara, S., De Gloria, A. (eds.) Applications in Electronics Pervading Industry, Environment and Society, pp. 280–285. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-95498-7_39
7. Dahiya, R., et al.: Large-area soft e-skin: the challenges beyond sensor designs. Proc. IEEE **107**(10), 2016–2033 (2019)
8. Alameh, M., Abbass, Y., Ibrahim, A., Moser, G., Valle, M.: Touch modality classification using recurrent neural networks. IEEE Sens. J. **21**(8), 9983–9993 (2021)
9. Osta, M., et al.: An energy efficient system for touch modality classification in electronic skin applications. In: 2019 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–4. IEEE (2019)
10. Silvera Tawil, D., Rye, D., Velonaki, M.: Interpretation of the modality of touch on an artificial arm covered with an EIT-based sensitive skin. Int. J. Robot. Res. **31**(13), 1627–1641 (2012)
11. Alameh, M., Ibrahim, A., Valle, M., Moser, G.: DCNN for tactile sensory data classification based on transfer learning. In: 2019 15th Conference on Ph.D Research in Microelectronics and Electronics (PRIME), pp. 237–240. IEEE (2019)

12. Jung, M.M., Poel, M., Poppe, R., Heylen, D.K.J.: Automatic recognition of touch gestures in the corpus of social touch. J. Multimodal User Interfaces **11**(1), 81–96 (2016). https://doi.org/10.1007/s12193-016-0232-9
13. Alameh, M., Abbass, Y., Ibrahim, A., Valle, M.: Smart tactile sensing systems based on embedded CNN implementations. Micromachines **11**(1), 103 (2020)
14. Gandarias, J.M., Gómez-de Gabriel, J.M., García-Cerezo, A.: Human and object recognition with a high-resolution tactile sensor. In: 2017 IEEE Sensors, pp. 1–3. IEEE (2017)
15. Gianoglio, C., Ragusa, E., Gastaldo, P., Valle, M.: A novel learning strategy for the trade-off between accuracy and computational cost: a touch modalities classification case study. IEEE Sens. J. **22**(1), 659–670 (2021)
16. Amin, Y., Gianoglio, C., Valle, M.: Computationally light algorithms for tactile sensing signals elaboration and classification. In: 2021 28th IEEE International Conference on Electronics, Circuits, and Systems (ICECS), pp. 1–6 (2021)
17. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
18. Kitronyx. https://www.kitronyx.com/store/p31/. Accessed 14 Apr 2022
19. Kitronyx. https://www.kitronyx.com/store/p68/Tinn_Force_Touch_Controller.html. Accessed 14 Apr 2022
20. Kitronyx. http://sites.kitronyx.com/wiki/applications/snowforce-3. Accessed 14 Apr 2022
21. Lele, P., Sinclair, D., Weddell, G.: The reaction time to touch. J. Physiol. **123**(1), 187 (1954)

Human Recognition for Resource-Constrained Mobile Robot Applied to Covid-19 Disinfection

Andrea Mattia Garavagno(✉), Daniele Leonardis, and Antonio Frisoli

Institute of Mechanical Intelligence, Scuola Superiore Sant'Anna of Pisa, PERCRO Lab, Via Alamanni, 13b, San Giuliano Terme, Ghezzano, 56010 Pisa, Italy
AndreaMattia.Garavagno@santannapisa.it

Abstract. The global COVID-19 pandemic has stimulated the use of disinfection robots: in September 2021, following a European Commission's action, 200 disinfection robots were delivered to European Hospitals. UV-C light is a common disinfection method, however, direct exposure to UV-C radiation is harmful and disinfection can be operated only in areas strictly forbidden to human personnel. We believe more advanced safety mechanisms are needed to increase the operational flexibility and safety level. We propose a safety mechanism based on vision and artificial intelligence, optimised for execution on mobile robot platforms. It analyses in real-time four video streaming and disables UV-C lamps when needed. Concerning other detection methods, it has a relatively wider and deeper range, and the capability to operate in a dynamic environment. We present the development of the method with a performance comparison of different implementation solutions, and an on-field evaluation through integration on a mobile disinfection robot.

Keywords: Covid-19 · UV-C based disinfection · Artificial intelligence · Human recognition

1 Introduction

The global COVID-19 pandemic has stimulated the production of disinfection robots by institutions and companies. In September 2021 the delivery of 200 disinfection robots produced by the UVD Robots company to European Hospitals was concluded [9]. It was part of the European Commission's action to help hospitals cope with the coronavirus.

Given the results, the Commission ordered further 100 robots, bringing the total donations to 300. This is the following of a trend that began in 2020 in China when the first disinfection robots were applied in the Smart Field Hospital in Wuhan [16].

The concept of automated disinfection without involving human operators looks interesting from the perspective of the hospital management, and not only.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 116–125, 2023.
https://doi.org/10.1007/978-3-031-16281-7_12

It can save lives by avoiding the cleaning personnel working in highly infected environments. At the same time, it can reduce costs by diminishing the required number of human operators. Consequently, these robots are being applied also in the disinfection of public spaces, such as the metro [17] and supermarkets [15].

The most commonly adopted robots, like the one from the UVD company [13,14], use UV-C light to disinfect surfaces. UV-C radiations alter DNA and RNA so that organisms cannot replicate. Others use also vapour and fogging systems that spray chemical disinfectants, such as ozone [10].

However, these robots are not ready to operate in a fully automated and flexible way. Because of the presence of UV-C lamps, strict safety mechanisms have to be applied, since UV-C radiation must not reach humans. In the end, there's a need for a human supervisor.

Evolving security mechanisms can reduce the risk of exposure to UV-C radiation due to supervisor errors.

2 Related Works

Today, autonomous disinfection robots seem to be the most effective way to reduce the risk of infections in public spaces such as hospitals, grocery stores, stations, etc. The most commonly adopted disinfection method uses UV-C light to disinfect surfaces, which has proven to be effective in reducing surface microbial contamination, cross contamination, and the spread of multi-drug resistant bacterial infections in hospitals [4]. Figure 1 shows an example of a UV-C-based disinfection robot.

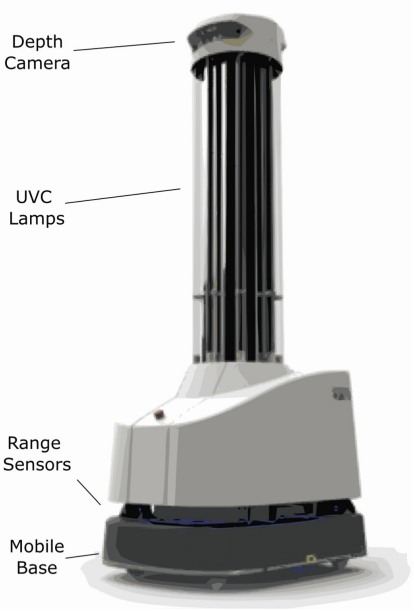


Fig. 1. An example of a UV-C-based disinfection robot.

UV-C radiations alter DNA and RNA so that organisms cannot replicate. Unfortunately, the majority of UV-C lamps emit small amounts of UV-B radiation which proved to be carcinogenic. Both high dose or prolonged low doses of radiation from some UV-C lamps can potentially contribute to skin cancer [11]. Nevertheless, even brief direct exposure of skin and eyes to UV-C radiation from UV-C lamps can cause painful eye injury and burn-like skin reactions [11]. Thus strict safety mechanisms are required to avoid radiation reaching humans.

Given that detecting human presence nearby is not an easy task, the majority of UV-C-based disinfection robots are designed to work in protected environments, where human is not present. Usually, the access to the space under

disinfection has to be restricted by obstacles, signals, or sensors at the entrances, which are used as safety measures. When one of them is removed the robot stops working since it assumes that a person entered the protected area. This procedure is the case of disinfection robots by UVD Robots [13,14].

Another common safety mechanism relies on the assumption that the protected environment is static. No motion is allowed. If something moving is detected, the robot has to stop working, waiting for the next activation by the supervisor. In this way, if a person accidentally enters the area under disinfection, its motion is detected and the lamps are shut down. This is the case of the disinfection robot proposed by Guettari et al. [1].

Motion detectors are not reliable solutions to detect human presence. A patient sleeping inside a hospital room doesn't produce enough movement to be detected. A careless check by the supervisor would expose the patient to the entire disinfection process.

Moreover, this mechanism strongly limits robots' cooperation capability. No other robot can work in the same area, since if something moves the disinfection is stopped.

Other disinfection robots proposed in scientific literature do not implement any human-detection safety mechanism [3].

The last type of safety mechanism encountered uses cameras and artificial vision to detect humans. This is the case of Ultrabot [2] by Perminov et al. and the PHS-M8 by Rayrobotics [12].

Ultrabot uses four cameras, placed in the front and the rear of the robot, to detect humans. However, no details about the algorithm operation and performance are provided, and the sensor placement does not seem to fully cover the operating area. No camera looks in the irradiation direction, so a human standing in front of the lamps would not be detected.

Instead, the PHS-M8 uses artificial intelligence to detect people by one video stream. The camera used cannot check the whole area surrounding the robot with its field of view. No details about the worst-case observed latency, the frame rate, and the accuracy are given.

3 Proposed Safety Mechanism

The proposed safety mechanism makes use of multiple camera sensors and Artificial Intelligence to detect the presence of humans. In particular, it implements convolutional neural networks (CNNs), to analyze in real-time the streaming video coming from four cameras mounted on board the disinfection robot. Figure 2 shows a schematic view of the hardware. This solution overcomes the limitations of the single-camera system while avoiding the issues related to motion detectors. Artificial vision algorithms can require in general relevant computational power to work in real-time: the main contribution of the presented work aims at developing a multiple-camera, real-time human detection system, capable of running on embedded processing units suitable for mobile robots applications.

The NVIDIA Jetson Nano has been chosen as the target computing platform. It is an embedded system equipped with an NVIDIA Maxwell graphic processing

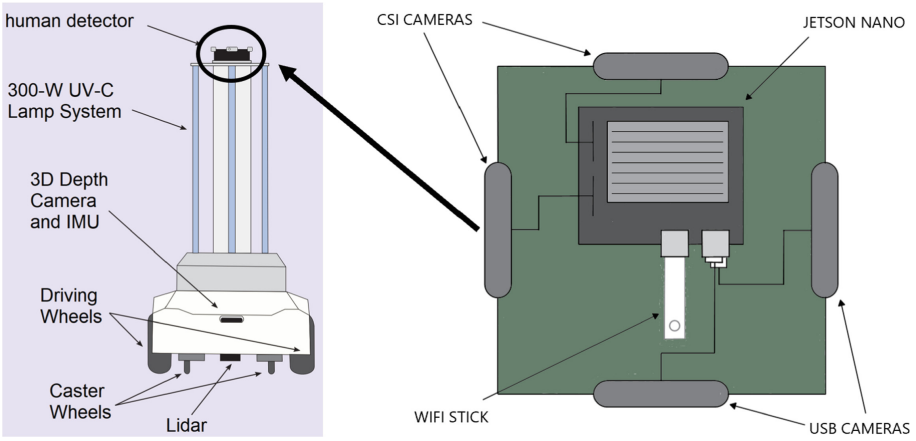


Fig. 2. A schematic view of the robotic mobile platform, on the left, alongside the human detector, on the right.

unit (GPU) with 128 NVIDIA CUDA®cores and 4 GiB of shared memory, able to run quite large CNNs, like VGG16, at 10 frames per second (FPS).

Two of the cameras used are Raspberry Pi Camera Module V2 connected through the two CSI connectors on the board, able to acquire at 1080p30, with an angle of view of $62.2 \times 48.8°$. The others two are Logitech C270 cameras able to acquire at 720p30, with a diagonal field of view of 55°, connected through two USB-2 connections.

The inference results are advertised through the robotic operating system (ROS) so that lamps are shut down whenever human presence is detected in at least one frame. Connectivity to other robot modules is provided by a WiFi USB stick.

The whole system is enclosed in an electronic box and placed on top of the robot, as shown in Fig. 2, to have a good point of view. The mobile disinfection robot used for the experimental evaluation was developed at the IIM Institute of the Scuola Superiore Sant'Anna of Pisa [3].

4 Design and Test of the Artificial Intelligence Application

The artificial intelligence application has been designed and trained thanks to the Keras API using Google's Colab free online service which offers limited GPU usage for research purposes. Then it has been exported in the ONNX data format and parsed by TensorRT to obtain an executable for NVIDIA GPUs. During the conversion, a cast to half-precision floating-point is performed. The obtained file has been executed directly on the onboard GPU thanks to the CUDA programming model obtaining the best performance.

The Common Object in COntext (COCO) data-set, 2017 version, is used to train and validate the application. It contains 118,287 training images, 64115

with humans and 54172 without, and 5,000 test images, 2693 with humans and 2307 without. Because of Google Colab access limits, only 12,000 images are used for training, 6,000 with humans and 6,000 without. Also, the test set has been balanced. 2,307 images are taken for both classes.

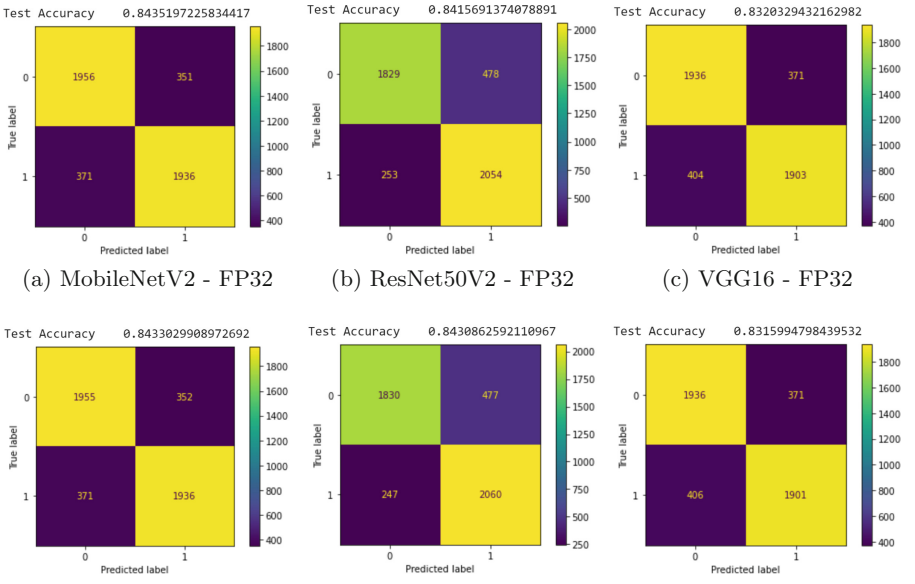
The test set is not used for training to better evaluate the generalization capability of the obtained application. Images with humans are labelled with 0 while the others with 1.

Transfer Learning is used to design the artificial intelligence application. Three different well-known pre-trained CNNs, offered by the Keras API, are tested: ResNet50V2, VGG16 and MobileNetV2. The weights come from training on ImageNet [6].

Only the feature extractor, with frozen weights, is taken. It is fed by a pre-processing pipeline. Then the extracted features are fed to a 2D Global Average Pooling layer to improve the model's generalization capability [7] and finally classified by a perceptron.

In the pre-processing pipeline, the input image is first resized to 224×224 resolution to save computations then min-max standardization is applied to improve the gradient descent convergence rate [8]. Adam [5] is the optimizer used.

The perceptron is trained for 10 epochs with a learning rate of 0.01 using binary cross-entropy as a loss function. If the output is less than or equal to 0.5 the image is classified as with human presence otherwise no. No data augmentation is applied.



(a) MobileNetV2 - FP32 (b) ResNet50V2 - FP32 (c) VGG16 - FP32

(d) MobileNetV2 - FP16 (e) ResNet50V2 - FP16 (f) VGG16 - FP16

Fig. 3. Confusion matrices - Keras (FP32) vs TensorRT (FP16).

Figure 3 shows the confusion matrices and the accuracy score computed onto the test set for each application. In the first line, the results obtained in Keras are shown. While in the second line the results computed on the Jetson Nano, using a half-precision floating-point are presented.

As shown from the presented graphs, the results are consistent. There are only a few changes in the confusion matrices due to the cast, which led MobileNetV2 and VGG16 to perform slightly worst. On the contrary, ResNet50V2 slightly improved its score.

Anyway, all the confusion matrices are well balanced except for the ResNet50V2 which shows a trend in detecting the human presence. This is might not be a problem, considering the aim is to build a safety mechanism. On the other hand, it might worsen the usability of the robot due to the higher rate of false positives.

Table 1. Accuracy, WCOET, and memory occupancy of each network executed on NVIDIA Jetson Nano.

	VGG16	MobileNetV2	ResNet50V2
Accuracy	0.8315	0.84330	0.84308
WCOET [s]	0.67	0.11	0.21
Footprint [MiB]	80.156	6.097	58.491

Looking at the Table 1 reporting the worst case observed latency (WCOET), the memory footprint and the accuracy score of each solution, the MobileNetV2 has been chosen as the best approach. It has the higher accuracy score and the lowest WCOET.

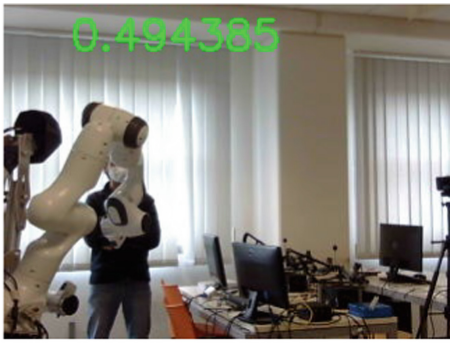
All the data have been measured on the Jetson Nano. The WCOET has been measured running one model at a time with 4 video streams active for 30 min each. It includes the time needed to gather the latest frame from each camera, the inference time, and the time to publish the result. The memory footprint has been taken by the file properties of the executable and the accuracy score is the one computed previously on the COCO 2017 test set.

5 On-Field Experiments

The proposed safety system based on the MobileNetV2 has been implemented and experimented on board the UV-C disinfection robot developed at the IIM Institute of the Scuola Superiore Sant'Anna. The robot was operated inside the laboratory's rooms during normal working activities to gather the frames processed by the system with the inference results overlaid on top. Green overlay numbers indicate a frame classified as human presence. Red numbers indicate the opposite.

During tests, activation of the UV-C lamps was inhibited by a safety switch, and the robot was teleoperated.

(a) Anthropomorphic head well classified



(b) Hidden human correctly classified

(c) Anthropomorphic robot

(d) High confidence

Fig. 4. Notable examples.

The gathered time-lapses evidenced two main behaviours: isolated glitches and gradual classification. Figures 5 and 6 show two examples of glitches in classification. Instead, gradual classification is shown in Fig. 7. As it can be seen in the images, occurred glitches appeared isolated. Given the relatively high frequency of the frame rate, these glitches can be avoided by implementation of a moving-window filter. In the real operating modality (with UV-C radiation activated) the filter would prevent fast power-off and power-on transients of the UV-C lamps. Anyway, using the whole dataset to train the networks would be a first better solution.

Finally, Fig. 4 shows some notable frames. In Fig. 4a there is an anthropomorphic head that has not been confused with a human; in Fig. 4b there is a human, partially hidden by a robotic arm, which is correctly classified despite the occlusion; in 4c there is an anthropomorphic robot correctly classified; and Fig. 4d shows a case in which the CNN is very confident about the classification.

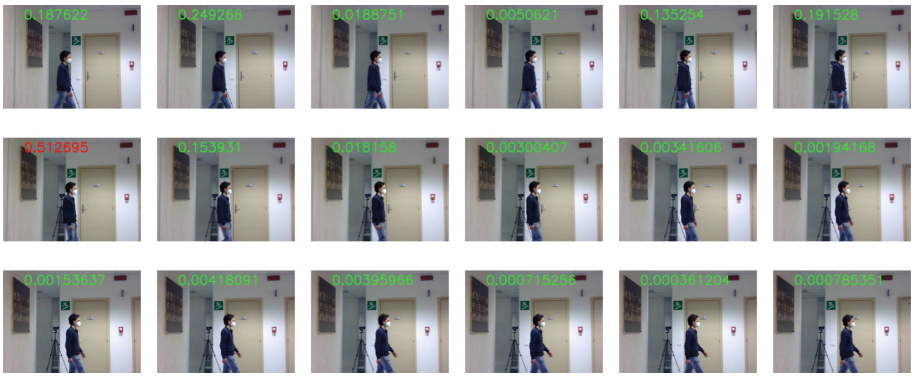


Fig. 5. Example of glitch with human in classification.

Fig. 6. Example of glitch without human in classification.

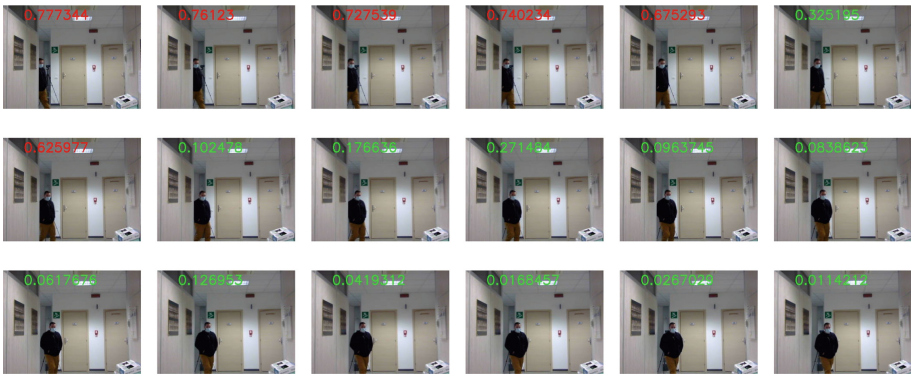


Fig. 7. Example of gradual classification of human.

6 Conclusion and Future Works

This paper presented a safety mechanism based on artificial vision that prevents UV-C radiations emitted by autonomous disinfection robots, from reaching humans. It enhances industrial and academic solutions by using artificial intelligence to analyze multiple video streams in real-time. The investigation was performed in order to develop and compare CNN-based algorithms suitable for execution on embedded processing units, typically available in mobile robot applications. The developed artificial intelligence application was validated on the 2017 version of the COCO test set, not used in training, and obtained an accuracy score of 84.3% with a balanced confusion matrix. The whole system has been implemented and experimented on a mobile disinfection robot to better understand its behaviour in a real case scenario. The evaluation was performed in an office unstructured environment: it confirmed results obtained during test, suggesting the method as a viable solution to increase safety levels on UV-C disinfection robots.

The safety mechanism can be further improved by using the whole COCO training set, applying data augmentation techniques, unfreezing and fine-tuning the last layers of the convolutional base, and connecting more cameras. Since Jetson Nano does not support 8-bit inference, no Quantization Aware Training can be applied to further reduce inference timings.

References

1. Guettari, M., Gharbi, I., Hamza, S.: UVC disinfection robot. Environ. Sci. Pollut. Res. **28**(30), 40394–40399 (2020). https://doi.org/10.1007/s11356-020-11184-2
2. Perminov, S., et al.: UltraBot: autonomous mobile robot for indoor UV-C disinfection. In: 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), pp. 2147–2152 (2021). https://doi.org/10.1109/CASE49439.2021.9551413
3. Tiseni, L., Chiaradia, D., Gabardi, M., Solazzi, M., Leonardis, D., Frisoli, A.: UV-C mobile robots with optimized path planning: algorithm design and on-field measurements to improve surface disinfection against SARS-CoV-2. IEEE Robot. Autom. Mag. **28**(1), 59–70 (2021). https://doi.org/10.1109/MRA.2020.3045069
4. Casini, B., et al.: Evaluation of an Ultraviolet C (UVC) light-emitting device for disinfection of high touch surfaces in hospital critical areas. Int. J. Environ. Res. Public Health **16**, 3572 (2019). https://doi.org/10.3390/ijerph16193572
5. Kingma, D.P., Jimmy, B.A.: Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, Conference Track Proceedings, 7–9 May 2015. arXiv preprint arXiv:1412.6980 (2014)
6. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
7. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
8. Shanker, M., Hu, M.Y., Hung, M.S.: Effect of data standardization on neural network training. Omega **24**(4), 385–397 (1996). https://doi.org/10.1016/0305-0483(96)00010-2

9. Coronavirus: 200th EU disinfection robot delivered to European hospital, a further 100 confirmed. https://digital-strategy.ec.europa.eu/en/news/coronavirus-200th-eu-disinfection-robot-delivered-european-hospital-further-100-confirmed
10. Disinfection Robot using Ozone. https://www.klainrobotics.com/wp-content/uploads/2020/03/PHS-Pro-Health-System-Brochure2.pdf
11. FDA - Ultraviolet (UV) Radiation. https://www.fda.gov/radiation-emitting-products/tanning/ultraviolet-uv-radiation
12. Rayrobotics. https://www.raybotics.it/#prodotti
13. UVD Robot Model B. https://uvd.blue-ocean-robotics.com/modelb
14. UVD Robot Model C. https://uvd.blue-ocean-robotics.com/robot-model-c
15. UVD Robots Move into the Supermarket Aisle. https://www.blue-ocean-robotics.com/news/uvd-robots-partners-with-occ
16. Autonomous Robots Are Helping Kill Coronavirus in Hospitals. https://spectrum.ieee.org/autonomous-robots-are-helping-kill-coronavirus-in-hospitals
17. Metro operator deploys cleaning robot to fight the coronavirus. https://www.railwaygazette.com/technology-data-and-business/metro-operator-deploys-cleaning-robot-to-fight-the-coronavirus/55995.article

Data-Driven Methods for Aviation Safety: From Data to Knowledge

Irene Buselli[1], Luca Oneto[2(✉)], Carlo Dambra[1], Christian Verdonk Gallego[3], and Miguel Garcia Martinez[1,2,3]

[1] ZenaByte s.r.l., Via Cesarea 2, 16121 Genova, Italy
{irene.buselli,carlo.dambra}@zenabyte.com
[2] University of Genoa, Via Opera Pia 11a, 16145 Genova, Italy
luca.oneto@unige.com
[3] Crida, Juan Ignacio Luca de Tena 14, 28027 Madrid, Spain
{ceverdonk,mgmartinez}@e-crida.enaire.es

Abstract. Demand upon the future Air Traffic Management (ATM) systems is expected to grow to possibly exceed available system capacity, pushing forward the need for automation and digitisation to maintain safety while increasing efficiency. This work focuses on a manifestation of ATM safety, the Loss of Separation (LoS), exploiting safety reports and ATM-system data (e.g., flights information, radar tracks, and Air Traffic Control events).

Current research on Data-Driven Models (DDMs) is rarely able to support safety practitioners in the process of investigation of an incident after it happened. Furthermore, integration between different sources of data (i.e., free-text reports and structured ATM data) is almost never exploited.

To fill these gaps, the authors propose (i) to automatically extract information from Safety Reports and (ii) to develop a DDM able to automatically assess if the Pilots or the Air Traffic Controller (ATCo) or both contributed to the incident, as soon as the LoS happens.

The LoSs' reported in the public database of the Comisión de Estudio y Análisis de Notificaciones de Incidentes de Tránsito Aéreo (CEANITA) support the authors' proposal.

Keywords: ATM · Safety · Digitisation · Data-Driven Models · Random Forests · Machine Learning · Loss of Separation · Safety reports

1 Introduction

In the last decades, air transportation has seen a considerable increase in demand, and pressure over Air Traffic Management (ATM) system is predicted to grow to possibly exceed the currently available system capacity [21]. The Single European Sky ATM Masterplan [22] defines the modernisation of the European ATM system as a crucial process to maintain safety while increasing

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 126–136, 2023.
https://doi.org/10.1007/978-3-031-16281-7_13

efficiency. A cornerstone of the Masterplan is to further deploy automation and digitisation tools, leading to an integration of human and technical systems [16].

This work is framed within the context of H2020 FARO project - saFety And Resilience guidelines for aviatiOn - which focuses on a manifestation of ATM safety, the loss of separation (LoS). There are two main sources of data which can inform about what happened during a LoS:

- the Safety Reports produced by states' Civil Aviation Authorities and ANSPs after investigating the safety-related events;
- the Automatic Safety Monitoring Tools (ASMTs), which allow the monitoring and recording of safety-related events. These tools are usually augmented with ATM system data, which gather surveillance (e.g., flight tracks) and operational data (e.g., ATC events) [4].

In particular, thanks to the new ASMTs, there are a number of safety-related occurrences - previously unnoticed by the old systems - which can now be identified. As a consequence, probably many more LoSs will need to be investigated and studied in the future. Since human review of incidents is an extensive process, the objective of this work is to provide the ability to partially estimate the results of these investigations timely (i.e., a few minutes after the LoS). This would facilitate the safety practitioners in prioritising the investigations and in understanding potential precursors of these LoS events.

To reach this objective, it is vital to connect the historical information included in the reports with the one contained in the structured data. Manual analysis of safety reports is complex and requires considerable resources, so the approach proposed in this work is twofold:

- First, a simple mining of free-text safety reports is implemented, with the purpose of identifying the information needed to connect reports to structured data;
- Then, an Automatic Contribution Assessment model was developed, able to leverage data to assess if the Pilots or the Air Traffic Controller (ATCo) or both contributed directly to the incident, almost immediately after the LoS and before investigation.

The LoS events considered in this study are those reported in the public database of the Comisión de Estudio y Análisis de Notificaciones de Incidentes de Tránsito Aéreo (CEANITA).

The exploitation of Data-Driven Models (DDMs) in the ATM domain is quite extensive. Indeed, research has focused on a number of different fields, such as taxi-out time prediction [12,17], trajectory prediction [1,25], air traffic flow extraction [6,26], and flight delay prediction [5,24]. In the safety domain, some relevant applications of DDMs are proposed in literature to predict safety events or performance [7,19], or to provide safety metrics [2] or accident precursors [13]. However, there are very few references aiming at supporting safety practitioners in facilitating the investigation of an incident after it happened but before it is reported [18]. To the best of the authors' knowledge, the approach proposed in this work is a pioneering one in this field.

2 Scope of the Work

The scope of this work is to develop an Automatic Contribution Assessment model able to leverage ATM-system data to assess whether the Pilots or the ATCos or both contributed to the incident. The model was able to assess contribution well before (i.e., 10 min after the incident) human evaluation (usually concluded even weeks after the incident).

3 Data Description

Two different data sources were exploited in this study: the CEANITA reports (see Sect. 3.1) and the structured data from ENAIRE-CRIDA data warehouse, containing contextual information about the LoSs together with ATC events (see Sect. 3.2).

3.1 CEANITA LoS Reports

The considered CEANITA LoS reports consist of 70 safety reports, written in Spanish and published by Spanish Safety Aviation Agency (AESA), which is the Spanish Civil Aviation Authority, under the commission of CEANITA, covering safety-related occurrences in the Spanish airspace between January 2018 and July 2019.

Each report is written as a free text and contains the following information:

- *Initial situation*: the initial location and condition of the aircraft involved in the LoS is described with text and images.
- *Communications and radar tracks*: the communications of interest between ATCos and pilots.
- *Conclusions*: the dynamic of the LoS based on the main actions performed by the involved human actors, the main causes, and Pilots and ATCo contribution (classified as direct, indirect or none).

3.2 ENAIRE-CRIDA Contextual Information

The structured contextual information was provided by ENAIRE-CRIDA. In particular, they provided high-granularity ATM data such as flight plans, flight tracks, and ATM-processed information about the Spanish airspace. More precisely, two main sources were exploited:

- flight tracks and related contextual flight information (e.g., type, speed, and heading) and
- ATC events of the interactions between ATCos and the Controller Working Position (CWP).

The complete list of features used in this work can be seen in Fig. 1.

4 Methods

This section presents the methods and tools exploited to achieve the scope of the work (see Sect. 2) leveraging the data described in Sect. 3. Data-driven predictive models are able to learn relations between inputs (e.g., ATC events) and outputs (e.g., incident direct contribution) based on a series of examples (i.e., historical data).

In this work, the main model exploited is Random Forests [3], a state-of-the-art solution in the field of Shallow Machine Learning algorithms. Even if, currently, Deep Learning approaches [9] were shown to outperform Shallow Learning models in many tasks, they require a huge amount of data to be trained, which was not available for this research.

Random Forests are one of the most effective approaches in the family of ensemble methods. It is a tree-based ensemble algorithm, combining bagging to random-subset feature selection. In bagging, each tree is independently constructed using a bootstrap sample of the dataset. Random Forests add a further layer of randomness to bagging, also changing how trees are constructed (the best split at each node of the tree is chosen among a subset of predictors randomly sampled at that node). Eventually, a simple majority vote is taken for prediction. Random Forests, while being not too influenced by their hyperparameters [15], still require to tune the number of trees (since the more trees the more accurate the model is, this number is chosen by trading off accuracy and computational complexity), and the number of predictors to be randomly sampled during trees construction.

So, as just described, the data-driven predictive models need to be tuned (by finding the optimal hyperparameters); however, at the same time, their performance needs to be estimated in a rigorous statistical way, in order to estimate their behaviour in production environment. Model Selection and Error Estimation deal exactly with this problem [14]. Resampling techniques like k-fold cross validation and non-parametric bootstrap are commonly exploited solutions, which work well in many situations [14]. These techniques rely on a simple idea: the original dataset is re-sampled once or more, without replacement, to build three independent datasets called learning, validation, and test set. The learning set is exploited to train the model. The validation set is exploited to find the optimal hyperparameters (namely the ones that lead to the optimal performance). The test set is exploited to estimate the performance of the final model (in this way, the test is independent from both the learning and the validation, so results are statically sound and no data snooping is allowed [27]). Performance measures strongly depend on the task to be solved. In this case, dealing with classification problems (see Sect. 5), Accuracy, Confusion Matrix, Area Under the Receiving Operating Characteristics (AUC), F1 score, Sensitivity, and Specificity are the most commonly used metrics [23].

Once the model is built and has been confirmed to be sufficiently effective, it can be of interest to investigate how this model is affected by the different input features [10,11]. This procedure is called Feature Ranking and allows the user to detect if the features are appropriately taken into account by the learned

models, according to their relevance from the perspective of the domain experts. In particular, Feature Ranking based on Random Forest via Mean Decrease in Accuracy (i.e., the importance of each feature is assessed by randomly permuting the values of the feature and measuring the resulting increase in error) is one of the most effective techniques [8,20].

5 Experimental Results

This section shows how the methods presented in Sect. 4 were exploited to achieve the scope of the work (see Sect. 2) demonstrating the effectiveness of the proposed approach on the data described in Sect. 3. Specifically, Sect. 5.1 presents the results of automatic information extraction from the CEANITA reports (necessary to connect them with the relative structured data), while Sect. 5.2 reports the performance of the data-driven model in estimating who directly contributed to the incident before the actual human evaluation.

5.1 Automatic Information Extraction from Free Text

The information extraction from CEANITA reports was an ancillary process aimed at retrieving two sets of features:

- The features necessary to connect each (anonymised) report with the relative structured data: date, time, and position.
- The contribution of Pilots and ATCo, classified as direct or none in both cases - please note that both Pilots and ATCo may have directly contributed to the incident. In particular, contribution was assessed as direct in the 36% of cases for Pilots and in the 72% of cases for ATCo's.

This task was performed through a rule-based procedure, based on the automated search of keywords, characters, and punctuation signs. The connection with the ENAIRE-CRIDA data was successful and enabled the subsequent development of the Automatic Contribution Assessment.

5.2 Automatic Contribution Assessment

After the preliminary extraction of information (Sect. 5.1), a DDM was exploited to assess agents' contribution before (i.e., 10 min after the incident) human evaluation (which is a post-operation activity) based on the automatic analysis of ATC events and other contextual data (i.e., radar tracks of the aircraft and flight information). Furthermore, the analysis shows that this predictive model actually captured meaningful relations and not just spurious correlations from the data (see Sect. 4).

Specifically, for each incident, the goal was to predict:

- the Pilots' contribution, i.e., classified as direct or not and
- the ATCo's contribution, i.e., classified as direct or not.

The prediction was based on a total of 19 features covering:

- the flight type,
- the flight rule at the moment of the incident,
- the flight level at the moment of the incident,
- the airspace class at the moment of the incident, and
- for each of the 15 classes of ATC events recorded from 30 min before to 10 min after the incident, their number of occurrences. Considering this time window is fundamental since the contributions of ATCo and Pilots depend both on what was done to prevent the potential LoS and on how it was managed when it became an actual LoS;

Different white-box and grey-box models were tested on this problem (i.e., Decision Trees, Logistic Regression, and Random Forests). The choice of not testing black-box models was due to the necessity of identifying, at least partially, the underlying process, in order to provide safety practitioners with potential precursors and to verify how the model is affected by the different features. In the end, a Random Forest model was selected (see Sect. 4), as it was proven to outperform the other ones.

The model was trained on the 70 incidents for which recorded ATC events were available (the number of trees was set to 1000 and the number of predictors to be randomly sampled during trees construction was searched in $\{5, 6, 7, 8, 9\}$ according to what was described in Sect. 4). Random Forests facilitate the generation of different optimal models changing the cut-off of the voting (i.e., how many trees need to agree to decide for a particular class). By doing so, it was possible to report different models, maximising respectively: the AUC, the F1 score, the Sensitivity, and the Specificity. Moreover, Random Forests provide the confidence of the prediction: this allows the user to trust the model only when its confidence is higher than a certain threshold.

Table 1 reports the confusion matrices of the developed predictive models (maximising AUC, F1 score, Sensitivity, and Specificity) for both ATCos' and Pilots' contributions.

Table 2, instead, reports the confusion matrices of the predictive models (maximising the AUC, since they appeared to be the most balanced ones) when predictions are considered only if their confidence is higher than 60% and 75%.

Table 1 shows that:

- When the AUC is maximised (i.e., assuming the user wants a balanced accuracy on both "Yes" and "No" classes), accuracy reaches ≈75% for Pilots contribution and ≈81% for ATCo; F1 score is ≈70% for Pilots and ≈86% for ATCo.
- When the F1 score is maximised (i.e., assuming the user wants to maximise the accuracy for the "Yes" class without too many false positives), accuracy reaches ≈73% for Pilots contribution and ≈85% for ATCo; F1 score is ≈70% for Pilots and ≈91% for ATCo.
- When the Sensitivity is maximised, (i.e., assuming the user wants to be as sure as possible that if the Pilots/ATCo contribute to the LoS the algorithm

Table 1. Confusion matrices of the developed predictive models of contribution based on the ATC events (maximising AUC, F1 score, Sensitivity, and Specificity) for both ATCo and Pilots contributions.

[Pilots Contribution (Maximising AUC)]

		Pred. No	Pred. Yes
Truth	No	46.0±0.3	18.3±0.3
	Yes	6.9±0.3	28.8±0.3

[ATCo Contribution (Maximising AUC)]

		Pred. No	Pred. Yes
Truth	No	20.8±0.3	4.9±0.3
	Yes	14.1±0.3	60.2±0.3

[Pilots Contribution (Maximising F1 score)]

		Pred. No	Pred. Yes
Truth	No	41.7±0.3	22.6±0.3
	Yes	4.0±0.3	31.7±0.3

[ATCo Contribution (Maximising F1 score)]

		Pred. No	Pred. Yes
Truth	No	12.1±0.3	13.6±0.3
	Yes	1.6±0.3	72.7±0.3

[Pilots Contribution (Maximising Sensitivity)]

		Pred. No	Pred. Yes
Truth	No	33.9±0.3	30.4±0.3
	Yes	0.1±0.2	35.6±0.2

[ATCo Contribution (Maximising Sensitivity)]

		Pred. No	Pred. Yes
Truth	No	11.9±0.3	13.8±0.3
	Yes	1.4±0.2	72.9±0.2

[Pilots Contribution (Maximising Specificity)]

		Pred. No	Pred. Yes
Truth	No	64.3±0.0	0.0±0.0
	Yes	20.3±0.2	15.4±0.2

[ATCo Contribution (Maximising Specificity)]

		Pred. No	Pred. Yes
Truth	No	24.9±0.2	0.8±0.2
	Yes	45.2±0.3	29.1±0.3

classifies it as "Yes") the level of sensitivity reached is ≈100% for Pilots, with ≈70% of accuracy, and ≈98% for ATCo, with ≈85% of accuracy; F1 score is ≈70% for Pilots and ≈91% for ATCos.
- When the Specificity is maximised (i.e., assuming the user wants to be as sure as possible that if the Pilots/ATCo are not responsible, the algorithm classifies it as "No") the level of specificity reached is ≈100% for Pilots, with ≈80% of accuracy, and ≈96% for ATCo, at the price of a low accuracy, ≈54%. F1 score is ≈60% for Pilots and ≈56% for ATCos.

Furthermore, Table 2 shows that:

- When just predictions with confidence ≥ 75% are considered, the accuracy reaches ≈97% for Pilots contribution and ≈94% for ATCo. With this threshold, only 43% of the predictions are trusted when assessing Pilots contribution and 44% when considering the ATCo.
- When, instead, the accepted confidence level is decreased from 75% to 60%, the accuracy reaches ≈86% for both Pilots and ACTo contributions. With this new confidence level, 62% of observations are classified when assessing Pilots contribution and 70% when considering ATCo.

Table 2. Confusion matrices of the developed predictive models based on the ATC events (maximising AUC) for both ATCo and Pilots contributions when predictions are trusted only if their confidence is higher than 60% and 75%.

[Pilots contribution (Confidence ≥60%)]

		Pred. No	Pred. Yes
Truth	No	58.1±0.3	9.3±0.3
Truth	Yes	4.7±0.3	27.9±0.3

[ATCo contribution (Confidence ≥60%)]

		Pred. No	Pred. Yes
Truth	No	20.4±0.3	6.1±0.3
Truth	Yes	8.2±0.3	65.3±0.3

[Pilots contribution (Confidence ≥75%)]

		Pred. No	Pred. Yes
Truth	No	60.0±0.0	00.0±0.0
Truth	Yes	3.3±0.1	36.7±0.1

[ATCo contribution (Confidence ≥75%)]

		Pred. No	Pred. Yes
Truth	No	29.0±0.1	3.2±0.1
Truth	Yes	3.2±0.2	64.6±0.2

Finally, the ranking of the features (see Sect. 4) produced by the Random Forest algorithm is presented in Fig. 1 to better understand what the predictive model actually learned from the data.

Figure 1 allows us to observe that, based on the experience of the domain experts, the models learned correctly the importance of features related to the separation responsibility, such as the Flight type, the Flight rules, or the Airspace Class. In addition, the models learned correctly the relevance of interactions between the ATC and the CWP, such as Radar Contact, ETO Over Fix or Action on Flight Level, in order to identify ATM contributions. These are promising results as the model presents room for improvement, such as the inclusion of more surveillance information or operational indicators such as traffic load.

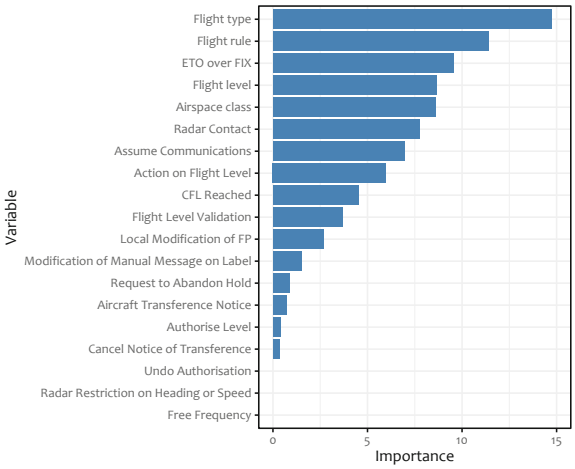


Fig. 1. Average variable importance ranking of the models (metric: mean decrease in accuracy).

6 Conclusions

The objective of this work was to facilitate the automatic extraction of meaningful and actionable information from LoS reports and to investigate how the information recorded by the systems can help estimating contribution assessment. For this purpose, the authors proposed a twofold approach based on (i) an automatic extraction of quantitative features from free text and (ii) an automatic contribution assessment model based solely on the information recorded by the systems and available a few minutes after the ASMTs' identification of the LoS. The approach was tested on the LoSs reported in the CEANITA public database and the related ATC events.

Different performance metrics were considered to evaluate the validity of the result. In particular, the results show that when only high-confidence predictions

are considered, the model output reaches approximately 97% of accuracy for pilots' contribution and 94% for ATCo.

Future work could validate these techniques on other databases of reports (e.g., UKAB AirProx Board, NTSF Board, etc.). Moreover, integrating other sources of structured data (e.g., about weather phenomena, STCA or TCAS activation, or traffic load) to develop richer models could lead to further insights in the estimation of contributors and precursors.

Acknowledgements. This project has received funding from the SESAR Joint Undertaking (JU) trough EU-H2020-ICT Project FARO - saFety And Resilience guidelines for aviatiOn (G.A. 892542). The dissemination reflects only the authors' view and the SJU is not responsible for any use that may be made of the information it contains.

References

1. Ayhan, S., Samet, H.: Aircraft trajectory prediction made easy with predictive analytics. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)
2. Bati, F., Withington, L.: Application of machine learning for aviation safety risk metric. In: IEEE/AIAA Digital Avionics Systems Conference (2019)
3. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
4. CANSO: Incidents investigation toolbox (2021). https://canso.fra1.digitaloceanspaces.com/uploads/2021/04/CANSO-Incidents-Investigation-Toolbox.pdf
5. Choi, S., Kim, Y.J., Briceno, S., Mavris, D.: Prediction of weather-induced airline delays based on machine learning algorithms. In: IEEE/AIAA Digital Avionics Systems Conference (2016)
6. Conde Rocha Murca, M., DeLaura, R., Hansman, R.J., Jordan, R., Reynolds, T., Balakrishnan, H.: Trajectory clustering and classification for characterization of air traffic flows. In: AIAA Aviation Technology, Integration, and Operations Conference (2016)
7. Di Gravio, G., Mancini, M., Patriarca, R., Costantino, F.: Overall safety performance of Air Traffic Management system: forecasting and monitoring. Saf. Sci. **72**, 351–362 (2015)
8. Genuer, R., Poggi, J.M., Tuleau-Malot, C.: Variable selection using random forests. Pattern Recogn. Lett. **31**(14), 2225–2236 (2010)
9. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep Learning. MIT Press Cambridge (2016)
10. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–42 (2018)
11. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**(Mar), 1157–1182 (2003)
12. Lee, H., Malik, W., Jung, Y.C.: Taxi-out time prediction for departures at Charlotte airport using machine learning techniques. In: AIAA Aviation Technology, Integration, and Operations Conference (2016)
13. Nazeri, Z., Barbara, D., De Jong, K., Donohue, G., Sherry, L.: Contrast-set mining of aircraft accidents and incidents. In: Perner, P. (ed.) ICDM 2008. LNCS (LNAI), vol. 5077, pp. 313–322. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-70720-2_24

14. Oneto, L.: Differential privacy theory. In: Model Selection and Error Estimation in a Nutshell. MOST, vol. 15, pp. 87–97. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-24359-3_9
15. Orlandi, I., Oneto, L., Anguita, D.: Random forests model selection. In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN) (2016)
16. Performance Review Commission, EUROCONTROL: Performance review report. An assessment of Air Traffic Management in Europe during the calendar year 2019 (2020). https://www.eurocontrol.int/sites/default/files/2020-06/eurocontrol-prr-2019.pdf
17. Ravizza, S., Chen, J., Atkin, J.A.D., Stewart, P., Burke, E.K.: Aircraft taxi time prediction: comparisons and insights. Appl. Soft Comput. **14**, 397–406 (2014)
18. Robinson, S.D., Irwin, W.J., Kelly, T.K., Wu, X.O.: Application of machine learning to mapping primary causal factors in self reported safety narratives. Saf. Sci. **75**, 118–129 (2015)
19. Rodríguez-Sanz, Á., Gómez, F., García, J.M.C., Meler, L.: Analysis of saturation at the airport-airspace integrated operations. In: USA/Europe Air Traffic Management Research and Development Seminar (2017)
20. Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008. LNCS (LNAI), vol. 5212, pp. 313–325. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87481-2_21
21. SESAR Joint Undertaking: European ATM master plan - executive view, 2015 edition (2015). https://www.sesarju.eu/node/2865
22. SESAR Joint Undertaking: European ATM master plan - executive view, 2020 edition (2020). https://op.europa.eu/en/publication-detail/-/publication/8afa1ad9-aac4-11ea-bb7a-01aa75ed71a1
23. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, Cambridge (2014)
24. Takeichi, N., Kaida, R., Shimomura, A., Yamauchi, T.: Prediction of delay due to air traffic control by machine learning. In: AIAA Modeling and Simulation Technologies Conference (2017)
25. Verdonk Gallego, C.E., Gómez Comendador, V.F., Amaro Carmona, M.A., Arnaldo Valdés, R.M., Séz Nieto, F.G., García Martínez, M.: A machine learning approach to air traffic interdependency modelling and its application to trajectory prediction. Transp. Res. Part C: Emerging Technol. **107**, 356–386 (2019)
26. Verdonk Gallego, C.E., Gómez Comendador, V.F., Saez Nieto, F.J., GarcíMartinez, M.: Discussion on density-based clustering methods applied for automated identification of airspace flows. In: IEEE/AIAA Digital Avionics Systems Conference (2018)
27. White, H.: A reality check for data snooping. Econometrica **68**(5), 1097–1126 (2000)

Design and Deployment of an Efficient Landing Pad Detector

Andrea Albanese[2(✉)], Tommaso Taccioli[1], Tommaso Apicella[1], Davide Brunelli[2], and Edoardo Ragusa[1]

[1] Electrical, Electronics and Telecommunication Engineering and Naval Architecture Department (DITEN), University of Genoa, Genoa, Italy
4398901@studenti.unige.it, tommaso.apicella@edu.unige.it, edoardo.ragusa@edug.unige.it

[2] Department of Industrial Engineering (DII), University of Trento, Trento, Italy
{andrea.albanese,davide.brunelli}@unitn.it

Abstract. Robust landing pad detection plays a major role in Autonomous Unmanned Aerial Vehicles (UAVs). This problem can be approached using deep neural networks for vision-based inference. However, the full integration of deep learning algorithms into the small UAVs is still challenging for their limited resources. This paper presents a landing pad detection pipeline based on a revisited version MobileNetV3-Small. The proposed architecture inherits robustness from the general-purpose version but limits the computational cost significantly thanks to a set of design criteria aimed to limit hardware requirements. Experimental results confirm that the proposed network compares favorably with a lightweight general-purpose object detector in terms of accuracy/computational cost trade-off. The system is also deployed on a commercial general-purpose microcomputer confirming that satisfactory performance can be obtained on general-purpose embedded architectures.

Keywords: CNNs · Embedded systems · Autonomous Unmanned Aerial Vehicles · Edge computing

1 Introduction

The successful application of Autonomous Unmanned Aerial Vehicles (UAVs) necessarily passes through the solution of complex problems of computer vision (CV). Although Deep learning (DL) proves to be the state-of-the-art for CV, resource-constrained devices struggle in supporting DL solutions in real-time. Eventually, this becomes a major limitation for the development of UAVs.

This paper presents a lightweight deep neural network (DNN) that can detect landing pads in aerial images. The proposed vision-based landing detector consists of a DNN based on a simplified version of MobileNetV3-Small [1]. As merit, the proposal limits the computational impact considering all the peculiarities of

A. Albanese and T. Taccioli—Equally contributed to the Manuscript.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 137–147, 2023.
https://doi.org/10.1007/978-3-031-16281-7_14

the problem under analysis. The intuition derived from the analysis of the problem was combined with a layer ablation study that confirmed the suitability of the proposed changes in the neural architecture.

As a major result, this solution enables autonomous control of small UAVs by processing the task directly in embedded systems mounted in the drones. The autonomous inference enables the UAV to detect a safe landing area and navigate accordingly to land on it.

As a proof of concept, the network was deployed in the Raspberry Pi 4 model B Single Board Computer (RPi) to evaluate its performance with an edge computing device. RPi sets a challenging benchmark because is not designed to host DL inference, but it is commonly used as a computing platform on small-size drones.

The approach was validated using a real-world dataset of aerial images. Results confirmed that the proposed network compares favorably with a recently proposed object detector, especially in terms of computing cost. In addition, the measure of energy consumption and the inference time confirm the suitability of the proposed approach.

The contribution of the paper can be summarized as follows:

- The paper presents a detailed analysis that leads toward the application of a deep learning pipeline for robust landing pad detection;
- The ablation study provides insight into the computer vision problem that can be employed in future studies;
- The deployment on a RPi confirms that the lightweight network can be supported by commercial devices suitable for the target application.

2 Related Works

Small UAVs may feature limited computational and energy resources, challenging autonomous navigation. So far, visual-based navigation is the most used configuration because low-power computer vision systems are becoming more and more accessible [2,3]. Researchers have recently proposed combining vision-based systems with Machine Learning (ML) algorithms to bring intelligence into UAV's missions. [4] combines traditional computer vision algorithm with DNN to develop a safe landing area detection system. This combination allows the preliminary process of the video to extract relevant features and then feed the DNN with the preprocessed video enhancing system efficiency. This is a successful approach which enhances the system's performance and energy efficiency. [5] proposes a two-stage training system to detect landing areas. A CNN trained over synthetic images is used in the first stage, then a custom Kalman Filter controller is used to have an accurate control during landing and approach the landing site. The second stage consists of a transfer learning approach that uses weights produced by the first stage CNN. With this two-stage system, it is possible to have a flexible model that can be employed in different scenarios. In general, ML algorithms can be used in mobile robotics applications adopting one of the two main approaches: cloud computing and edge computing.

Cloud computing consists of transmitting sensed data to a server with near-unlimited computational resources to compute the prediction using a ML model and send the result back. However, data transmission introduces a considerable latency and energy consumption. For instance, [6] presents an object detection system for UAVs by using Region-based Convolutional Neural Networks (R-CNN) and the cloud computing approach. "Dronetrack" [7] is another system developed for UAVs that use cloud-based real-time object tracking. This paper revealed that fast connections such as optic fiber lead to real-time execution. These results suggest that running ML applications with a cloud computing approach is feasible. However, these systems are network-dependent. Thus, if it is considered an application that can be scaled in different scenarios where it is not always possible to have a stable wireless connection, the cloud computing approach is not the best choice.

The edge computing approach overcomes the cloud limitations leading to better execution performance. For this reason the most disparate applications were refined to be hosted in edge devices, including but not limited to semi-autonomous prosthesis [8], pest detection [9,10], visual sentiment analysis [11]. For instance, [12] presents a system for perception, guidance, and navigation for racing drones. This system uses DL techniques to estimate the racing gate centers and assist drones in navigating through gates. It uses NVIDIA Jetson TX2 as a computer on-board responsible for computing the prediction in real-time, achieving a frame rate of 28.95 FPS. The same board is used in [13] for enabling real-time object detection in a UAV warning system. It exploits the state-of-the-art YOLO V2 [14] for recognizing dangerous products with visual aerial inspection directly on the drone board. Platforms such as GPUs and SBCs can accommodate complex models requiring a high memory footprint, while MCUs can host hundreds of KBs. Thus, the choice of the target platform is conditioned to the model that has to host. For instance, [15] use Raspberry SBCs as target platform to enable autonomous navigation on UAVs. The used CNNs are tailored to produce fast results with the computational power of Raspberry and with a low-power profile. On the other hand, if it is necessary to use huge ML models with resource-constrained platforms (e.g., MCUs), optimization techniques become necessary to fit ML models in the target platform.

Aside from platforms development, a surge of dedicated DL architectures were proposed in recent years to enable edge computing. The most used optimization techniques for predefined architectures are pruning [16], weights compression [17], quantization [18], and low-rank approximation [19]. However, the starting architecture significantly bias memory footprint, run-time memory requirements, and the number of floating point operations performed during inference phase (FLOPs). In recent years, researchers have proposed new operations that decrease computation requirements: depthwise convolution (DWConv) [20], mobile inverted bottleneck (MBConv) [21], efficient activation [22]. The latest research trends for lightweights architecture definition were focused on knowledge distillation [23] and Neural Architecture Search (NAS) [24]. In practice, most of the latest general purpose CNN for computer vision such as MobileNetV3 [25]

and EfficientNet [26] were obtained using NAS-based procedure. However, NAS-based approaches induce a computational overload in the training phase, which requires one to rely on dedicated computing equipment. In addition, the constraints introduced by a device such as a low-cost microcontroller would heavily impact the design of the eventual loss function, thus affecting the optimization phase.

3 Proposal: Design of a DNN for Landing Detection

MobileNetV3-Small architecture was proposed as a lightweight general-purpose deep network [1]. The network is built on top of a core building block called Bottleneck. The original paper uses the backbone network to target two different applications: object detection and semantic segmentation. Semantic segmentation provides a bit-wise classification of the pixels in the form of masks. Instead, object detection produces a number of bounding boxes, each one associated with a class and confidence. In some cases, DNNs for semantic segmentation feature fewer parameters than object detection alternatives. In our case, the two solutions presented in MobileNetV3 paper featured 0.47 M parameters for segmentation and 1.77 M parameters for the detector. Given the smaller footprint, we propose to use a customized version of the semantic segmentation architecture for landing pad detection. In the last part of the section, we present the low-cost strategy that converts the mask into the coordinates of the landing pad.

Even being one of the smallest DNNs for image classification MobileNetV3 is largely oversized for the target problem. This architecture was designed to extract high-level semantic features; meanwhile, the target problem involves the detection of target objects with simple geometric patterns. For this reason, large parts of the network are useless and inflate the compute cost. It is well-known that the higher layer of the networks extracts high-level semantic features. In the original implementation, the 9th bottleneck block and the last 2D convolution fed the segmentation head [1]. We propose the 2nd and 5th bottleneck blocks as inputs to the segmentation head. This means that the layers after the 5th bottleneck block have been excluded from the architecture, thus improving execution speed and reducing the total number of parameters. Among the candidates, the 2nd and 5th bottleneck blocks were selected to trade-off generalization capabilities and compute cost because the size of the output tensors elaborated by these two blocks matched the input size required by the segmentation head.

In addition, the architecture has been tailored to handle small resolution images. The original design of MobileNet-V3 supports input with 1024×2048 pixels resolution. High-resolution images are not required in landing pad detection, since it could be sufficient to distinguish the landing pad from the background, and introduce an increment in resources utilization. Furthermore, the original architecture uses stride different from 1 in the first level of the architecture that can negatively affect the landing pad's detection when distant. We tackled this issue by setting the stride of the first convolution layer and second bottleneck layer from 2 to 1. The stride acts as a compression factor on the image size.

Table 1. Proposed Backbone. KEY - Operator: building block; Exp size: expansion factor; Out: number of output filters; SE: Squeeze-And-Excite; NL: nonlinear activation; RE: ReLU; HS: hard swish; s: stride.

Input	Operator	Expsize	Out	SE	NL	s
$320^2 \times 3$	conv2d, 3×3	–	16	–	HS	1
$320^2 \times 16$	bneck, 3×3	16	16	v(1)	RE	2
$160^2 \times 16$	bneck, 3×3	72	24	–	RE	1
$160^2 \times 24$	bneck, 3×3	88	24	–	RE	1
$160^2 \times 24$	bneck, 5×5	96	40	v(2)	HS	2
$80^2 \times 40$	bneck, 5×5	240	40	v(3)	HS	1

Eventually, the new version of the network envisions an input size of $[320 \times 320]$. The final output of the designed DNN is a heatmap H of size 160×160 with a probability value that corresponds to the probability of the pixel belonging to a landing pad. Table 1 summarizes the architecture of the proposed backbone network highlighting the most import parameter on the columns.

Squeeze-And-Excite (SE) layer plays a major role in the original definition of MobileNetV3, both in terms of generalization performance and compute cost. On the one hand, these blocks contain fully connected layers that contribute significantly to the final number of parameters. On the other hand, SE effectively propagates the information toward different layers of the network. Given that we used a small portion of the original network (around half of the original depth), we can speculate that the contribution of SE layers will be negligible. Eventually, this operation reduced the memory requirements significantly of the network with a negligible impact on the performance of the target problem. An ablation study will confirm this claim in the experimental setup.

The lightweight pipeline in Fig. 1 performs all the tasks that convert the output of the segmentation network into the 2D coordinates of the landing pad. First, a threshold operation filter out the pixel with a probability lower than a reference value. Then, the well-known algorithm described in [27] groups the pixel with a significant probability in clusters, as shown in the figure. Finally, the cluster having the largest perimeter is selected as the predicted box.

4 Experiments

4.1 Generalization Performance Validation

A dataset composed of 13 videos of landing pads containing a total of 17,720 frames has been collected. The videos are collected from three different heights: approximately from 3, 4 and 6 m. In each frame one out of three landing pad types are present. Figure 2 reports three examples of images from the dataset with additional zero-padding. 5,542 frames from 5 of these videos have been employed as training sets. The remaining images from the same videos have

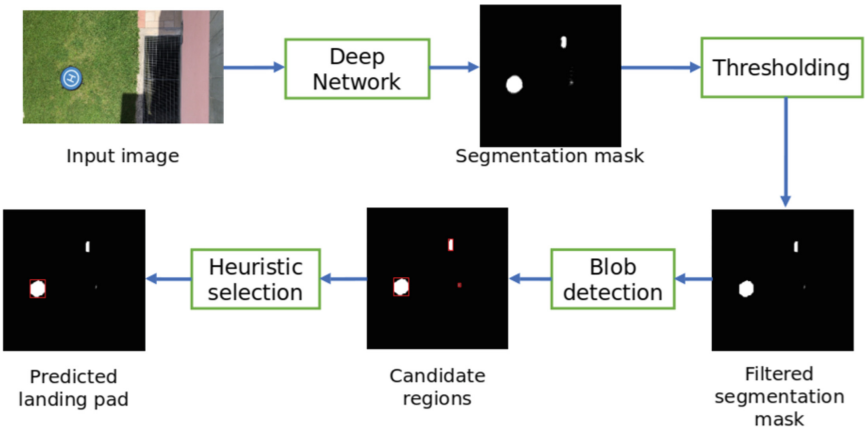


Fig. 1. Pipeline from the RGB input to the landing pad detection.

been then discarded, leaving 11,300 images for the test. The experiments involved two distinct test sets. A first set, called test1, contains all the testing images. A subsection of this testing set, called test2, corresponding to 3,328 images belonging to 2 videos never included in training operations, has been employed as a more challenging test set. We obtained the masks using a weak labeling strategy. A MobileNetV3-SSD (Small) pretrained on COCO dataset for object detection has been fine-tuned on a subset of the dataset that has been manually labeled. This network has been used to annotate the remaining images. The eventual bounding boxes have been further refined using a thresholding strategy based on the color of the landing pad. It should be noted that this step was a semi-manual annotation and cannot be included in a fully autonomous pipeline.

Visual inspection from two human users confirmed that the prediction obtained were sufficiently overlapped with the landing pad. Eventually, the MobileNetV3 - SSD sets the comparison in the experimental setup.

Fig. 2. Example of landing pad images contained in the dataset.

An ablation study confirmed the hypothesis about the contribution of the SE layers. The network was trained for 10 epochs, with batch size 4, learning rate 0.0001. Table 2 summarizes the results of the ablation study, showing the performance of the network in terms of Intersection over Union (IoU). The IoU, a.k.a. Jaccard index, is a measure of how much the predicted bounding box (A) is comparable with respect to the ground truth (B) following the relationship: $IoU = \frac{|A \cap B|}{|A \cup B|}$. We compared 5 configurations, Full refers to the network as presented in Table 1, *SE(n) excluded* identifies the network without the n-th SE layer. The second and third columns show the number of parameters of the complete architecture and the number of floating-point operations for one inference phase, respectively. Column IoU_{tot} reports the average IoU value measured between the labels and the weak labels. Errors report the number of images where the IoU value was 0. Finally, column IoU_{clean} reports the average IoU only for images where the measured IoU was greater than 0. The performance was measured using the whole test set, i.e., test1.

Table 2. Ablation study results. KEY - Params: Number of parameters; FLOPs: Number of floating-point operations; G: 10^9; IoU_{tot}: average Intersection over Union; IoU_{clean}: average of Intersection over Union greater than 0.

Architecture	Params	FLOPs (G)	IoU_{tot}	Errors	IoU_{clean}
Full	195,460	1.295	71.6%	1,093 (9.7%)	79.3%
SE(1) excl.	194,916	1.294	77.5%	31 (0.3%)	77.7%
SE(2) excl.	176,836	1.293	76.5%	31 (0.3%)	76.7%
SE(3) excl.	79,780	1.291	69.0%	403 (3.6%)	71.6%
SE(1,2,3) excl.	60,612	1.289	72.8%	530 (4.7%)	76.4%

The results highlight that SE blocks play a major role in the number of parameters. In particular, SE(3) has a huge impact on the number of parameters. The number of operations is heavily affected by these layers, but the relative contribution is smaller because dense layers impact weights but not significantly on tensor size propagation. In terms of generalization error, we can note that SE blocks affect overall performance but do not always positively affect the network. The solution without SE blocks proves more accurate than the version without the third SE when images with IoU 0 are excluded. The larger number of errors for the Full architecture was because, in some cases, masks were significantly noisy. A visual inspection confirmed that the prediction marked as an error was correct in a few cases.

Table 3 summarizes the comparison with the MobileNetV3-SSD (V3-SSD), i.e. the solution that would have been naturally employed from [1]. The table reports number of parameters (in millions) and FLOPs measured in Giga (G) followed by the average IoU computed over the challenging test set, i.e., test2. IoU columns report, in addition, between brackets, the average IoU considering only images where IoU was greater than 0.

Table 3. Comparison with state-of-the-art model for lightweight object detection. KEY - Params: Number of parameters; M: 10^6; FLOPs: number of G: 109; IoU: average Intersection over Union.

Architecture	Params (M)	FLOPs (G)	IoU
V3-SSD	2.761	3.758	69.3%
Proposal	0.061	1.289	61.7% (63.7%)

The results highlight that the proposed approach includes 45 times fewer parameters and requires around one-third of operations with respect to the original object detector. When analyzing the IoU, one must consider that those analyses in biased towards the SSD. In fact, this model has been used to build most of the labels. Accordingly, the 8% gap in terms of the IoU can be ascribed to noise introduced by the weak labeling process.

4.2 Model Optimization

The developed system for landing pad detection was deployed on the RPi board. The system was characterized by considering energy consumption, frame rate, and memory usage for processing one frame by averaging the measured values. Moreover, the system performance is compared among four model configurations taking into account the architecture presented in Sect. 3 with INT8, FP16, and FP32 bit representation, along with MobileNetV2 (MobV2) based setup used in [15] for landing pad detection with FP32 bit representation. The measurement campaign was conducted by processing about 700 frames for each configuration in a static scene. Specifically, the system was mounted in order to view and process a landing pad at a distance of three meters continuously. Moreover, a General Purpose Input Output (GPIO) was configured to give an impulse of 100 ms at the end of each frame to synchronize the processed frames with the relative current consumption. An oscilloscope was used to monitor the current consumption as well as the GPIO with a sampling frequency of 50 kHz. The measurements of current consumption relative to the processing time were analyzed to find the average time and the average current consumption. The results of the overall system characterization are summarized in Table 4.

As shown in Table 4, the 32 bit configuration is the most efficient and with the best frame rate. The 8 bit configuration, indeed, shows the worst performance because it is not supported by the RPi's instruction set. Moreover, MobileNetV2 needs an additional piece of pipeline to solve the problem of landing pad detection, thus confirming the efficiency of the proposed solution. The overall system characterization for the best configuration (FP32) reveals that the average processing of one frame takes 388.05 ms and consumes 1.80 J. Moreover, it uses only 2.27% of the available RAM (i.e., 4 GB), achieving 2.58 Frames Per Second (considering a the average time to perform the processing), enough to compute a landing trajectory and correct eventual drifts. The presented system runs onboard on UAVs with the edge computing approach. Considering a small-size

Table 4. Comparison between the proposed model and MobileNetV2 [15] in terms of energy consumption, execution time, measurement standard deviation and memory usage, varying the quantization. KEY - Config: quantization used; INT8: 8-bit integers; FP16: Floating-point 16 bits; FP32: Floating-point 32 bits; MobV2: MobileNetV2 quantized FP32; std: Energy standard deviation; Time: average time; FPS: Frames Per Second considering the average time; RAM: Random-access memory.

Config	Energy (J)	std	Time (ms)	FPS	RAM	Disk usage
INT8	2.22	0.0051	523.51	1.91	97.05 MB	108.17 kB
FP16	1.83	0.012	390.31	2.56	92.69 MB	132.98 kB
FP32	**1.80**	0.012	388.05	**2.58**	92.82 MB	766.31 kB
MobV2	1.80	0.0097	478.80	2.09	223.01 MB	26.70 MB

UAV, it is equipped with a 1500 mAh battery operating at 14.8 V which ensures a flight time of 15 min. With the proposed system operating onboard, the drone's available energy and flight time are reduced by only 5%. This means that bringing intelligence to UAV systems does not compromise the available energy, which is a precious resource in such systems.

5 Conclusions

This paper provides a detailed analysis that leads toward the application of a deep learning pipeline for robust landing pad detection. The analysis is carried out through an ablation study that considers insights of the analysed computer vision problem. The eventual deployment on a RPi confirms that the lightweight network can be supported by commercial devices suitable for the target application. As future work, the integration of the RPi module with hardware accelerators e.g. Intel Neural Compute Stick, will provide additional boost to the performance of the network and additional clues concerning the real-time performance of the pipeline. Moreover, the employment of NAS techniques will lead to the development of an even more efficient DL model, achieving better performance, extending this solution in other UAVs autonomous navigation applications.

References

1. Howard, A., et al.: Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324 (2019)
2. Courbon, J., Mezouar, Y., Guénard, N., Martinet, P.: Vision-based navigation of unmanned aerial vehicles. Control Eng. Pract. **18**(7), 789–799 (2010)
3. Fauadi, M.H.F.M., et al.: Intelligent vision-based navigation system for mobile robot: a technological review. Period. Eng. Nat. Sci. **6**(2), 47–57 (2018)
4. Paszkuta, M., et al.: UAV on-board emergency safe landing spot detection system combining classical and deep learning-based segmentation methods. In: Nguyen, N.T., Chittayasothorn, S., Niyato, D., Trawiński, B. (eds.) ACIIDS 2021. LNCS (LNAI), vol. 12672, pp. 467–478. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-73280-6_37

5. Mathur, P., Jangir, Y., Goveas, N.: A generalized Kalman filter augmented deep-learning based approach for autonomous landing in MAVs. In: International Symposium of Asian Control Association on Intelligent Robotics and Industrial Automation (IRIA), vol. 2021, pp. 1–6. IEEE (2021)
6. Lee, J., Wang, J., Crandall, D., Šabanović, S., Fox, G.: Real-time, cloud-based object detection for unmanned aerial vehicles. In: 2017 First IEEE International Conference on Robotic Computing (IRC), pp. 36–43. IEEE (2017)
7. Koubâa, A., Qureshi, B.: DroneTrack: cloud-based real-time object tracking using unmanned aerial vehicles over the internet. IEEE Access **6**, 13810–13824 (2018)
8. Ragusa, E., Gianoglio, C., Dosen, S., Gastaldo, P.: Hardware-aware affordance detection for application in portable embedded systems. IEEE Access **9**, 123178–123193 (2021)
9. Segalla, A., Fiacco, G., Tramarin, L., Nardello, M., Brunelli, D.: Neural networks for pest detection in precision agriculture. In: IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), vol. 2020, pp. 7–12. IEEE (2020)
10. Albanese, A., Nardello, M., Brunelli, D.: Automated pest detection with DNN on the edge for precision agriculture. IEEE J. Emerging Sel. Top. Circ. Syst. **11**(3), 458–467 (2021)
11. Ragusa, E., Apicella, T., Gianoglio, C., Zunino, R., Gastaldo, P.: Design and deployment of an image polarity detector with visual attention. Cogn. Comput. **14**(1), 261–273 (2022)
12. Jung, S., Hwang, S., Shin, H., Shim, D.H.: Perception, guidance, and navigation for indoor autonomous drone racing using deep learning. IEEE Robot. Autom. Lett. **3**(3), 2539–2544 (2018)
13. Tijtgat, N., Van Ranst, W., Goedeme, T., Volckaert, B., De Turck, F.: Embedded real-time object detection for a UAV warning system. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 2110–2118 (2017)
14. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
15. Albanese, A., Nardello, M., Brunelli, D.: Low-power deep learning edge computing platform for resource constrained lightweight compact UAVs. Sustain. Comput. Inform. Syst. **34**, 100725 (2022)
16. Ye, S., et al.: Progressive DNN compression: a key to achieve ultra-high weight pruning and quantization rates using ADMM. arXiv preprint arXiv:1903.09769 (2019)
17. Yang, E.-H., Amer, H., Jiang, Y.: Compression helps deep learning in image classification. Entropy **23**(7), 881 (2021)
18. Park, E., Ahn, J., Yoo, S.: Weighted-entropy-based quantization for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5456–5464 (2017)
19. Kim, Y.-D., Park, E., Yoo, S., Choi, T., Yang, L., Shin, D.: Compression of deep convolutional neural networks for fast and low power mobile applications. arXiv preprint arXiv:1511.06530 (2015)
20. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
21. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)

22. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions. arXiv preprint arXiv:1710.05941 (2017)
23. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: a survey. Int. J. Comput. Vis. **129**(6), 1789–1819 (2021)
24. Zhang, L.L., Yang, Y., Jiang, Y., Zhu, W., Liu, Y.: Fast hardware-aware neural architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 692–693 (2020)
25. Howard, A., et al.: Searching for MobileNetV3. CoRR arXiv:1905.02244 (2019)
26. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
27. Suzuki, S., et al.: Topological structural analysis of digitized binary images by border following. Comput. Vis. Graph. Image Process. **30**(1), 32–46 (1985)

Towards a Trade-off Between Accuracy and Computational Cost for Embedded Systems: A Tactile Sensing System for Object Classification

Youssef Amin(✉), Christian Gianoglio, and Maurizio Valle

Department of Electrical, Electronic, Telecommunication Engineering and Naval Architecture - DITEN, University of Genoa, Genoa, Italy
{youssef.amin,christian.gianoglio}@edu.unige.it, maurizio.valle@unige.it

Abstract. The deployment of the inference phase in self–standing systems, which have resource–constrained embedded units, is faced with many challenges considering computational cost of the elaboration unit. Therefore, we propose using a learning strategy based on a loss function that leads to finding the best configuration of the prediction model balancing the generalization performance and the computational cost of the whole elaboration system. We validate our proposal by integrating a tactile sensing system on a Baxter robot to collect and classify data from five daily–life objects using four different algorithms. Results show that the best performance, when the computational cost is not relevant, is achieved by the fully–connected neural network using 16 features, while, when the computational cost matters, the loss function showed that the kernel SVM with 4 features has the best performance.

Keywords: Tactile sensing systems · Object classification · Artificial intelligence for embedded devices · Computational load

1 Introduction

In recent decades, many research works have been focused on deriving object properties from tactile signals – using Machine Learning (ML) – to differentiate between different daily–life objects. Several studies have adopted the human approach for tactile object identification, the 'exploratory procedures' (EPs) [1], where typical movement patterns are carried out to collect information about objects. However, such approach demands a wide range of tactile data processing, thus it increases the work complexity. Furthermore, new technologies have emerged such as tactile sensing systems that consist of tactile sensors and embedded electronics. These embedded units of such systems are required to perform signal processing, feature extraction, and prediction. Usually, ML algorithms are utilized to perform the classification. However, the execution of all these tasks can be computationally very expensive for resource – constrained

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 148–159, 2023.
https://doi.org/10.1007/978-3-031-16281-7_15

devices. Knowing that each of these tasks affects the predictor performance and the computational cost, it is important to find the best configuration (i.e. signal processing techniques, extracted features, model architecture, etc.,) for the elaboration unit, that trades–off between the computational cost and the predictor performance, before deploying the elaboration unit into a prior designed embedded system.

In this paper we present the implementation of a learning strategy based on the evaluation of a loss function [2], in order to find the best configurations of the elaboration unit stages, addressing the classification of 5 objects based on tactile data. The learning strategy provides a balance between the computational cost of the whole elaboration unit (i.e. in terms of FLOPs) and the generalization performance of the predictors. In this regard, we apply an automated technique to remove unneeded information from a single unplanned grasping action (inspired from the concept of haptic glance found in humans [3]) and extract features, to reduce the computational cost on the elaboration unit. Moreover, for prediction, we implement ML algorithms that have proven to be suitable for resources–constrained devices [4,5].

The remainder of the paper is: Section 2 presents related works, Sect. 3 describes our proposal, Sect. 4 explains the experimental setup and the workflow, Sect. 5 presents the results, eventually Sect. 6 concludes the paper.

2 Related Works

Before a few decades, The awareness of robotic systems to their surrounding environment was confined by visual exploration, which helps to understand the workplace for better manipulation. However, to acquire certain object properties (i.e. hardness, texture, weight, etc.,) physical interaction with objects is required, and in some cases manipulation is needed. Furthermore, ML and deep learning algorithms have been utilized to extract high–level information from tactile data. Some of these studies have aimed to classify texture [6], stiffness [7,8], or objects recognition [9–11]. The authors in [7,12] integrated a tactile sensing system into a Baxter robot to collect data from three objects of different hardness. They used for the classification a single layer feed-forward neural networks (SLFNNs) that proved to be effective in terms of generalization performance and the computational cost of the predictor. Similarly in [13], the authors proposed a methodology to employ and then optimize classifiers of varying complexity using a principle optimization algorithm. The later algorithm, allowed to automatically configure the coarse–to fine–networks, as it achieves a trade–off between energy consumption and accuracy for the inference. As a result for that, the optimized corase–to–train network reduced the overall energy for image classification problem without any loss in accuracy. The authors in [9] utilized a four–fingered hand that compresses 241 tactile skin sensors distributed over the hand, equipped on TWENDY–ONE robot. The data was collected by grasping 20 different daily–life objects, of different shapes and hardness. For object recognition, a deep learning neural network approach was employed along with a

denoising auto–encoder (DAE) and a dropout. In addition, two feature learning techniques: the self–organizing maps and the principal component analysis were applied. For the comparison, a single layer neural network was implemented and the recognition rates for different sensor modalities were tested. whereas in [10], the authors evaluated the performance of different neural network–based models, namely Convolutional Neural Networks (CNN) and Long–Short Term Memory (LSTM), on an object recognition task for 9–classes and compare the results on tactile grasping data collected using two different fingertip tactile robot sensors. Experimental results show that the LSTM–based model outperforms the CNN models. They also present a method to extend the available data for training, leading to performance improvement. Authors in [11] proposed a hybrid methodology for performing tactile classification and feature extraction, having a row of TakkTile tactile pressure sensors, integrated into an under–actuated robotic hand. The setup was used to collect data while performing a single unplanned grasp. However, exploratory motions were not applied. Features were extracted from the sensors readings and the actuator position, at three instances of the grasping process, and employed for object classification using the random forest method. This approach provided a system with low computation and complexity overheads for haptic sensing applications. On the other hand, the authors proposed in [8] a semi–supervised generative adversarial network (GAN) for hardness detection to mitigate the expensive process of data labeling. Unsupervised training of GAN with a large number of unlabeled samples provides the architecture and initial parameter values for the hardness recognition network (HRN), which is trained with manually labeled samples corresponding to each hardness level. The hardness detection result can be determined online by importing the tactile data collected by the robotic forearm into the trained HRN.

In these works, the authors did not evaluate the computational cost of their models nether their elaboration unit, except in [7,13]. However, it is crucial to keep track of the computational cost along with the generalization accuracy knowing that high model accuracy can be obtained at the expense of higher computations. In this context, we compute the FLOPs for all the operations starting from signal processing (i.e. filtering of the tactile signal) until the prediction of object. We also employ a loss function in the learning strategy to obtain the best model configuration balancing between the accuracy and the computational cost for the whole elaboration unit.

3 Proposal

A tactile sensing system is made of a sensing patch and an elaboration unit. The elaboration unit includes different stages: signal processing, feature extraction, and prediction. All these stages require a certain amount of power consumption which may form a problem in resource–constrained devices. On the other hand, having a good model with a high generalization performance sometimes come with many expenses on the computational cost. Therefore, balancing the trade–off between generalization performance and computational cost becomes

crucial. We propose using a learning strategy based on a loss function that balances the generalization performance and the computational cost, keeping into consideration all the stages of the elaboration unit.

3.1 Loss Function

In [2], the authors proposed a loss function that during the training phase is capable of selecting the best model in terms of the trade–off between the predictor accuracy and its computational cost. Procedure 1 in [2] depicts the learning strategy that we adopt in this paper as well. It consists of a data–driven approach based on the out–of–sample technique [14] where the original dataset is split into training and validation sets. The training set is used to train the models on all the possible configurations of the hyper–parameters, solving (2). The validation set is then employed to find the best model minimizing the loss function, solving (1). The loss function consists of two terms:

$$i^* = \underset{i}{\operatorname{argmin}}\, \hat{L}_m(f^*_{n,i}) + \theta R_H(f^*_{n,i}) \tag{1}$$

where

$$f^*_{n,i} = \underset{f \in \mathcal{F}_i}{\operatorname{argmin}}\, \hat{L}_n(f) + \lambda R(f). \tag{2}$$

In (2), $\mathcal{F}_i$ represents one of the models indexed by the hyper–parameters, $\hat{L}_n(f)$ is the empirical risk on the training set, and $\lambda R(f)$ describes the regularizer with its weight. In (1), $\hat{L}_m$ denotes the empirical risk on the validation set, while the second term θR_H is the hardware constraint weighted by a parameter that balances the trade–off between accuracy and computational. Adopting a value of $\theta = 1$ means giving the same importance to the hardware constraint and the accuracy, while with a value greater than 1 the loss function promotes lighter model (e.g. model with lower size, lower number of parameters, lower FLOPs, etc.) with respect of achieving the best accuracy. Therefore, Procedure 1 is employed taking into account different combinations of the elaboration unit stages. In the following, we describe in detail the stages of the elaboration unit and their computational cost in terms of FLOPs.

3.2 Signal Conditioning

The tactile sensing system employed for this study is the same used in [7], which consists of a low–cost piezoelectric sensing array and a low–power interface electronics (IE) that has a low–current input analog–to–digital converters and is able to acquire data from multiple input channels (i.e. 32). We employ a method to reduce the number of samples acquired from each channel, by introducing a threshold that triggers a real–time extraction of Ns number samples during a grasp. Ns represents the length of the generated signal. Finally, we filter the generated signal $\tilde{\boldsymbol{x}}$ using a two–samples Moving Average filter (MA): $y[i] = \alpha\tilde{x}[i] + (1-\alpha)y[i-1]$; both α and Ns are user defined parameters. Based on [15], the computational cost of filtering with MA is $(3Ns + 5)$ FLOPs.

3.3 Feature Extraction

Similarly to [7], we extract as features the statistical moments from the filtered signals $\boldsymbol{y}$. They proved to be efficient from the computational cost side and led to a high accuracy during the prediction stage. Table 1 reports the number of FLOPs for each feature. We considered that is possible to employ partial results of a feature to compute the others. As an example, the mean and standard deviation of a signal can be used in the computation of the skewness, thus we save FLOPs.

Table 1. Extracted features and FLOPs count

Extracted features	FLOPs count
Mean	$Ns + 1$
STD	$3Ns + 5$
Skewness	$2Ns + 7$
Kurtosis	$2Ns + 5$

3.4 Predictors

We present four different ML algorithms to classify the five daily objects. One of the algorithms is based on the support vector machine (i.e. SVM), which aims to classify different classes by finding the hyper–planes that maximize the margin between two classes. In this work, a Gaussian kernel (i.e., the radial basis function RBF) is used with the SVM (K–SVM). The hyper–planes of SVM classifier are represented by the following function for a two–class problem: $f(\boldsymbol{z}) = \sum_{i=1}^{nSV} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{z}) + b$; where $\boldsymbol{z}$, nSV, α_i, y_i, $\boldsymbol{x}_i$, K, and b denote the tested datum, the number of support vectors, the coefficient of the support vector, the corresponding label, the training datum that lies on the support vector, the kernel, and the bias, respectively. More Information about SVM are in [16].

The other three algorithms are neural networks made of one hidden layer, thus recognized as single–layer feed–forward neural networks (SLFNNs). The SLFNNs are known to be able to provide high generalization performance in classification while maintaining the computational cost low. The general prediction function of a SLFNN for a two–class problem is: $f(\boldsymbol{z}) = \sum_{i=1}^{N} \beta_i \phi(\boldsymbol{z} \cdot \boldsymbol{\omega}_i + b_i)$; where $\boldsymbol{z}$ the tested datum, N is the number of hidden neurons, β_i the weights between the hidden and output layers, ϕ the activation function, $\boldsymbol{\omega}$ the weights between the input and hidden layers, and b_i the neuron bias. The first SLFNN is the fully–connected neural network that employs ReLU activation functions (R–FC), and is trained by the backpropagation technique. In contrast, the second and third SLFNNs are based on the Extreme Learning Machine (ELM) paradigm [17]. In ELM, parameters $\boldsymbol{w}, b$ are set randomly, and training demands solving a Regularized Least Square (RLS) problem to learn the beta connections. For the first ELM network, we assigned a ReLU activation function (R–ELM). While

for the second network, we adopted a hard limit activation function (H–ELM) and set the $\boldsymbol{w}$ weights as a power of two due to its effectiveness on low power and low cost devices [4,18].

Table 2 shows the FLOPs count of each of the four algorithms, where nf, nc, N, and nSV correspond to the number of features, number of classes, number of neurons, and number of support vectors. In case of a two–class classification problem $nc = 1$.

Table 2. SLFNNs and FLOPs count

SLFNNs	FLOPs Count
H–ELM	$N(2nf + 1) + nc * N$
R–ELM	$N(2nf + 2) + 2nc * N$
R–FC	$N(2nf + 2) + 2nc * N$
K–SVM	$\sum_{i=1}^{nc} nSV_i * (3nf + 10) + 1$

4 Experimental Setup

To evaluate the performance of different model architectures to classify grasped objects –by evaluating the loss function (1)– we integrated our tactile sensing system on the gripper of Baxter arm. The system comprises a P(VDF–TrFE) piezoelectric sensing patch and IE, installed on one side of the gripper, to acquire tactile data while running grasping experiments on five daily–life objects.

4.1 System Setup

Shown in Fig. 1 the system workflow. Similar to the previous work in [7], we choose the same tactile sensing patch, which has high sensitivity over a wide frequency bandwidth (0.5 Hz–1 kHz). This patch contains a matrix of eight piezoelectric sensors ($4 \times 2 = 8$ sensors) uniformly distributed, forming a rectangular sensing area of $2.1 \times 1.1\,\mathrm{cm}^2$. As well, it has a spatial resolution of 1 cm center–to–center pitch and a sensor diameter of 2 mm. Before doing any experiment, the patch is shielded –using a special conductive tape– to carry noises (i.e. external charges) to the ground. In addition, a substrate and a thin protective layer is added to the bottom and top side of the patch, respectively. we use the IE to acquire tactile data from sensors. The IE contains an ARM–cortex M0 micro–controller that leads to sample the signals on 32 channels with a frequency of 2KSamples/sec. To integrate our sensing patch into the Baxter robot, we designed a 3D printed gripper fingers that perfectly fit the sensing patch (see Fig. 2a). On the other hand, we fasten the low–power IE to the end of the Baxter arm using an elastic fabric strap. The IE is set to transmit continuously filtered tactile data to the host PC through a USB connection. The data collection process is controlled using a LabVIEW GUI, which we designed for reading,

visualizing, and saving the tactile data from eight sensors. Eventually, automatic extraction of grasp peak from each tactile signal, followed by features extraction, then the training of models is done offline using the PC (Fig. 1).

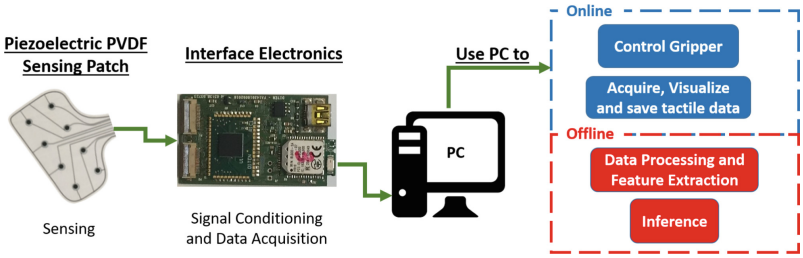


Fig. 1. System workflow.

4.2 Objects

For the experiments, we select a set of five different 'daily life' objects shown in Fig. 2b. These objects have three types of surfaces: flat (e.g. one side of the eyeglasses case and the cubic box), circular (e.g. the cola can and the tennis ball), and curved (e.g. the shampoo bottle and the other side of the eyeglasses case) surfaces. The objects are different in shape, materials, and size. But, they all have a good impact resistance (i.e. high stiffness) and are big enough to be in contact with many sensors when grasped.

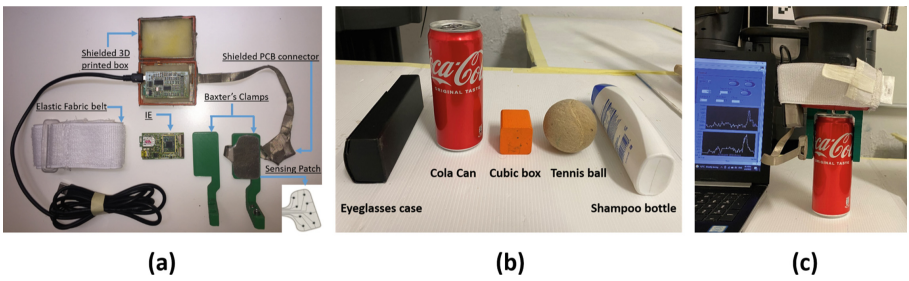


Fig. 2. a) System setup. b) Set of objects used. c) Example of object grasping.

4.3 Data Collection

Data was collected by grasping each of the five objects 400 times, at different positions, using the Baxter gripper as shown in Fig. 2c. Our experiments do not involve any post–grasp manipulation; therefore, the arm position is fixed. Before the experiments, we determine the closure position of the gripper for each object,

which corresponds to the gripper position when the touch occurs. We set the gripper velocity and force to $v = 5\,\text{cm/s}$ and $F = 0.03 * Fmax$ ($Fmax = 35\,\text{N}$). The chosen objects are hard enough to avoid any deformation when grasped. During each trial, the gripper closes –at a constant velocity and force–to the predefined position, and remains closed for 1.2 s; then, it opens for two seconds. The trials are repeated with no sensors and/or gripper feedback (open–loop control). Grasping control is done by the robotic operating system (ROS). The IE continuously acquires the tactile data from all sensors simultaneously at a high sampling frequency (2 KSps) and transmits them to the PC. According to Sect. 3.2, the signals are pre–processed before extracting the features. Using Matlab, we automatically extract 150 samples from the grasp peak by taking 25 samples before the sample that has the minimum value and 125 samples after. This process is done for all the channels simultaneously, and for all the trials.

Eventually, we build six datasets based on the combination of the number of sensors (i.e., excluding four sensors or using all of them) and the extracted features. Each dataset contains 2000 data, i.e. 400 samples for each class. Table 3 gives the details of the six datasets. Each column corresponds to a dataset, the first row reports the name of the datasets, while the second, third, fourth, and the fifth rows report the number of sensors used to build the datasets, the number of features extracted from each sensor, the type of feature (i.e. Mean, standard deviation = STD, kurtosis = Kur, and skewness = Skew), and the total number of features, respectively. The six datasets correspond to six different feature extraction stages. The loss function will be evaluated to find the best combination of features extraction stage and predictor.

Table 3. Generated datasets

Datasets names	4S1F	4S2F	4S4F	8S1F	8S2F	8S4F
Number of sensors	4	4	4	8	8	8
Number of features per sensor	1	2	4	1	2	4
Type of features	Mean	Mean STD	Mean, STD Skew, Kur	Mean	Mean STD	Mean, STD Skew, Kur
Total num of features	4	8	16	8	16	32

4.4 Training Strategy

Four algorithms are built to classify the five objects, therefore the SLFNNs have five output neurons while the K–SVM presents five linear separators by applying the 'one–vs–one' approach. In the SLFNNs, the Softmax activation function assigns the label to the neuron that has the highest probability. In the K–SVM, the label of the tested datum was assigned by evaluating its position with respect to all the separators. All the algorithms were trained over the six datasets previously described. The SLFNNs hyper–parameters are defined as follows: the hidden neurons $N = \{50, 100, 200, 300\}$, the regularization term L2

$\lambda = \{10^i,\ i = -4, -3, ..., 5\}$, random weights and biases between $[-1, 1]$ (for H–ELM they were represented as the power of 2 [4,18]); while for the K–SVM: the RBF kernels parameter $\gamma = \{10^i,\ i = -4, -3, ..., 5\}$, and same regularization term λ as for SLFNNs. For the FC, the number of epochs, batch size, patience on the validation accuracy for the early stop criterion, and the learning rate using the Adam optimizer were fixed to 200, 64, 20, and 0.01, respectively. The loss function (1) is evaluated by setting the parameter θ as $\theta = \{0, 0.5, 1, 2, 5, 10\}$. In this way, $\theta = 0$ means that only the accuracy is relevant during the evaluation, while the greater the θ the higher the relevance of having a light model with a low computational cost. For all classifiers, we apply a stratified K–folds cross–validation with $K = 10$: the datasets are divided into 10 folds which, in turn, 9 contribute to form the training set and 1 the test set. Therefore, each algorithm is trained 10 times on each dataset (i.e., 10 runs). The stratification leads to having the same proportion of data for each class in the folds with respect to the original dataset. In each run, the training data are randomly split in training/validation sets using the 75% of data as training and 25% as validation. The validation set is employed to evaluate the loss function for each θ value. Eventually, the test set is used to assess the generalization performance of the best models in terms of accuracy.

5 Results

Through the evaluation of the loss function, it is possible for each algorithm to find the best set of hyper–parameters and the best features set that leads to the best compromise between generalization accuracy and computational cost measured as the number of FLOPs, for all the values of the θ parameter. The FLOPs are considered along all the elaboration stages and normalized between 0 and 1. The number of FLOPs and generalization accuracies are averaged along the 10 folds, for each pair features set–algorithms and for each value of θ.

Procedure 1 in [?] is evaluated on each algorithm to find the best combination of features set and predictor hyper–parameters, for all the values of θ. Then, we compute the loss function (1) for each θ on the test data. Results are reported in Fig. 3. In the figure, the loss values are multiplied by 100 for a better description. For each theta, there are four colored bars that correspond to the following algorithms: K–SVM, R–FC, R–ELM, and H–ELM. For $\theta = 0$ (meaning that the computational cost is not relevant), the best loss function value is obtained by R–FC (99.6), followed by R–ELM (98.0), K–SVM (97.6), and finally H–ELM (97.2). In this case, the score corresponds to the test accuracy. However, for $\theta > 0$, the loss function decreases, and different results are achieved as the K–SVM outperforms the others. Between $\theta = 0.5$ and $\theta = 10$ the loss function decreases slightly in K–SVM ($\approx$5%), and significantly in R–FC ($\approx$20.3%), R–ELM ($\approx$20.5%), and H–ELM ($\approx$24%). As a result, the best loss function is always achieved by the K–SVM. For all $\theta > 0$, the H–ELM has the lowest values; which is normal because it requires much more neurons than R–ELM and R–FC to attain similar generalization performances in terms of accuracies [7].

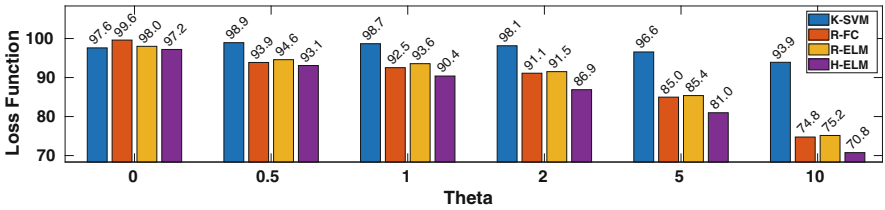


Fig. 3. The loss function of the four algorithms, at different θ weights.

Table 4. Best configurations showing the number of sensors (NS), extracted features, average FLOPs, and average generalization accuracy.

Algorithms	θ	NS	Features	Av. FLOPs	Av. accuracy STD
K–SVM	0	8	M&STD	9855	97.60 ±3.38
	0.5	4	M	2818	99.21 ±1.69
	1	4	M	2818	99.21 ±1.69
	2	4	M	2818	99.21 ±1.69
	5	4	M	2818	99.21 ±1.69
	10	4	M	2818	99.21 ±1.69
R–FC	0	8	M&STD	9300	99.60 ±1.27
	0.5	4	M	3296	95.20 ±3.16
	1	4	M	3296	95.20 ±3.16
	2	4	M	3156	95.20 ±3.16
	5	4	M	3156	95.20 ±3.16
	10	4	M	3156	95.20 ±3.16
R–ELM	0	8	M	5770	98.01 ±2.83
	0.5	4	M	3156	95.60 ±3.51
	1	4	M	3156	95.60 ±3.51
	2	4	M	3156	95.60 ±3.51
	5	4	M	3156	95.60 ±3.51
	10	4	M	3156	95.60 ±3.51
H–ELM	0	8	M&STD	9620	97.20 ±2.70
	0.5	4	M	3646	95.20 ±4.14
	1	4	M	3506	94.00 ±5.42
	2	4	M	3226	91.60 ±6.39
	5	4	M	3156	91.21 ±7.01
	10	4	M	3156	91.21 ±7.01

Table 4 presents the best configurations chosen by the evaluation of the loss function, showing the number of sensors and extracted features, based on each algorithm and θ value. In the table, the columns from the left to the right reports,

respectively, the following: algorithms, θ, number of sensors, extracted features, average FLOPs, and the average accuracy on the test set. For $\theta > 0$, we obtained the same dataset (i.e. 4F1S) for all algorithms; this dataset utilizes the minimal number of sensors (i.e. 4) and type of features (i.e. the mean M). Whereas, for $\theta = 0$, the best configuration for R–ELM (i.e. 8 sensors, one type of features) is different from the other algorithms (having 4 sensors and 2 types of features: M and STD). Moreover, for all algorithms, we observe that: the bigger the θ lower the averaged computational cost and accuracy, except in the case of K–SVM; nevertheless, it gives the highest accuracies with the lowest computational costs for $\theta > 0$. And, the R–FC attains the best accuracy for $\theta = 0$. Therefore, by employing the loss function in the learning strategy for each algorithm, not only did we find the configurations which provide the best trade–off between the generalization accuracy and number of FLOPs, but also the best predictor (K–SVM) for object classification when the computational cost is relevant.

6 Conclusion

This work presents a learning strategy based on a loss function to find the best configurations of algorithm, when balancing between the predictor accuracy and the computational cost of the whole elaboration unit (i.e. signal processing, feature extraction, and prediction). A parameter in the loss function is tuned to weight the importance of the computational cost. The loss function is employed for the classification of five objects, measuring the computational cost in terms of FLOPs. Results show that the best performance, when the computational cost is not relevant, is achieved by the fully–connected neural network with the ReLu activation function using 16 features while, when the computational cost matters, the loss function chooses the kernel SVM of 4 features as the best configuration. In the future work the model with the best trade-off can be chosen to be deployed on a resource-constrained devices.

Acknowledgment. The authors acknowledge financial support from TACTIle feedback enriched virtual interaction through virtual realITY and beyond (Tactility) project: EU H2020, Topic ICT-25-2018-2020, RIA, Proposal ID 856718.

References

1. Lederman, S.J., Klatzky, R.L.: Hand movements: a window into haptic object recognition. Cogn. Psychol. **19**(3), 342–368 (1987)
2. Gianoglio, C., Ragusa, E., Gastaldo, P., Valle, M.: A novel learning strategy for the trade-off between accuracy and computational cost: a touch modalities classification case study. IEEE Sens. J. **22**(1), 659–670 (2021)
3. Klatzky, R.L., Lederman, S.J.: Identifying objects from a haptic glance. Percept. Psychophysics **57**(8), 1111–1123 (1995)
4. Ragusa, E., Gianoglio, C., Zunino, R., Gastaldo, P.: A design strategy for the efficient implementation of random basis neural networks on resource-constrained devices. Neural Process. Lett. **51**, 1–19 (2019)

5. Ragusa, E., Gianoglio, C., Zunino, R., Gastaldo, P.: Random-based networks with dropout for embedded systems. Neural Comput. Appl. **33**(12), 6511–6526 (2020). https://doi.org/10.1007/s00521-020-05414-4
6. Chun, S., et al.: An artificial neural tactile sensing system. Nature Electron. **4**(6), 429–438 (2021). https://doi.org/10.1038/s41928-021-00585-x
7. Amin, Y., Gianoglio, C., Valle, M.: Computationally light algorithms for tactile sensing signals elaboration and classification. In: 2021 28th IEEE International Conference on Electronics, Circuits, and Systems (ICECS), pp. 1–6 (2021)
8. Qian, X., et al.: Hardness recognition of robotic forearm based on semi-supervised generative adversarial networks. Frontiers Neurorobot. **13**, 73 (2019)
9. Schmitz, A., Bansho, Y., Noda, K., Iwata, H., Ogata, T., Sugano, S.: Tactile object recognition using deep learning and dropout. In: 2014 IEEE-RAS International Conference on Humanoid Robots, pp. 1044–1050 (2014)
10. Bottcher, W., Machado, P., Lama, N., McGinnity, T.: Object recognition for robotics from tactile time series data utilising different neural network architectures. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2021)
11. Spiers, A.J., Liarokapis, M.V., Calli, B., Dollar, A.M.: Single-grasp object classification and feature extraction with simple robot hands and tactile sensors. IEEE Trans. Haptics **9**(2), 207–220 (2016)
12. Amin, Y., Gianoglio, C., Valle, M.: A novel tactile sensing system for robotic tactile perception of object properties. In: AISEM Annual Conference on Sensors and Microsystems. pp. 182–187. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-08136-1_28
13. Jayakodi, N.K., Chatterjee, A., Choi, W., Doppa, J.R., Pande, P.P.: Trading-off accuracy and energy of deep inference on embedded systems: A co-design approach. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **37**(11), 2881–2893 (2018)
14. Anguita, D., Ghio, A., Oneto, L., Ridella, S.: In-sample and out-of-sample model selection and error estimation for support vector machines. IEEE Trans. Neural Netw. Learn. Syst. **23**(9), 1390–1406 (2012)
15. Thant, H.A., San, K.M., Tun, K.M.L., Naing, T.T., Thein, N.: Mobile agents based load balancing method for parallel applications. In: 6th Asia-Pacific Symposium on Information and Telecommunication Technologies, pp. 77–82. IEEE (2005)
16. Noble, W.S.: What is a support vector machine? Nat. Biotechnol. **24**(12), 1565–1567 (2006)
17. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. Neurocomputing **70**(1–3), 489–501 (2006)
18. Ragusa, E., Gianoglio, C., Gastaldo, P., Zunino, R.: A digital implementation of extreme learning machines for resource-constrained devices. IEEE Trans. Circuits Syst. II Express Briefs **65**(8), 1104–1108 (2018)

An Optimized Heart Rate Detection System Based on Low-Power Microcontroller Platforms for Biosignal Processing

Benedetta Mazzoni[1(✉)], Giuseppe Tagliavini[1], Luca Benini[1,2], and Simone Benatti[3]

[1] University of Bologna, Bologna, Italy
b.mazzoni@unibo.it
[2] Swiss Federal Institute of Technology Zürich, Zürich, Switzerland
[3] University of Modena and Reggio Emilia, Modena, Italy

Abstract. The real-time detection of the R peaks of the ECG signal is crucial to provide information on cardiac functionality, and several strategies have been presented in the past. In this work, we adapt the classical Pan and Tompkins (PT) algorithm for efficient execution on low-power microcontroller (MCU) platforms to design a full-fledged heart rate detection system. We target a commercial MCU based on ARM Cortex-M4 and an ultra-low-power solution based on the RISC-V PULP platform. Experimental results show that our approach achieves an accuracy above 99.5%, comparable to the state-of-the-art solutions, and an energy efficiency that is one order of magnitude better than other software solutions.

Keywords: ECG · R peak · Heart rate · Low- power biosignal processing

1 Introduction and Related Work

The growing trend of small form factor devices is pushing the development of wearables, driven by systems such as health patches and trackers [1,2]. In the fitness and healthcare area, these systems facilitate remote and continuous monitoring of wellness conditions providing the extraction of physiological parameters from the analysis of biosignals [3]. However, the main weakness of wearable sensor nodes is the request for high computationally-demanding tasks at high energy efficiency to improve the battery life-time.

Several works exploit digital platforms capable to execute digital signal processing (DSP) to achieve ultra-low power (ULP) consumption [4,5]. In this context, the designers typically adapt optimization strategies to reduce the algorithm complexity and find the best trade-off between reliability and low power consumption. Among the biopotentials that can be acquired with real-time low

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 160–170, 2023.
https://doi.org/10.1007/978-3-031-16281-7_16

power devices [6], heart activity parameters are the most used to detect and monitor acute severe conditions. Analysing the QRS complex and detecting R peaks is crucial for providing cardiac functionality information. Several strategies have been presented in the literature, using well-established signal processing techniques. Park et al. [7] propose a technique based on a wavelet transform (WT) coupled with the Shannon energy envelope method in addition to a moving average filter and a squaring operation for the preprocessing step. This method achieves accuracy over 99%. However, the algorithm is computationally intensive, and it is not suitable for real-time execution on an ultra-low-power embedded device. Martinez et al. [8] adopt the phasor transform. This approach converts each ECG sample into a complex number maintaining the phase and the root mean square values to enhance the wave variations to distinguish them from each other. Also in this case, the overall accuracy is higher than 99%. However, the analysis excludes five records from the MIT-BIH Database because of the low-quality acquisition of highly-variable signals or noise distortion.

A widely explored family of approaches for ECG signal analysis includes *slope-based methods*. In Tekeste et al. [9], the authors optimize peak detection by providing a hardware unit to approximate the computation of the signal derivative. The power consumption is only 3.9 nW at an operating frequency of 3 kHz implemented in 65 nm technology. Nevertheless, they do not consider the contribution to the power consumption of the additional computations that are strictly required by a real-life scenario. In our work, we use microcontroller-class devices that can perform pre-processing, peak detection, and subsequent computations. De Giovanni et al. [10] propose a software-based methodology that can be considered the current state-of-the-art. Their algorithm implements a Bayesian filter, normalization, and a clustering technique to optimize the R peak detection. The authors test the system on a biosignal dataset where sudden event changes occur, such as during intense physical exercise [11]. These physical conditions reduce the robustness of the traditional algorithms affecting their reliability. Hence, they propose an accurate adaptive design for low-power platforms. However, the authors do not consider the aspects related to real-time signal acquisition. They use an existing system (BIOPAC) that requires a 9 V battery and is not energy efficient. Furthermore, the peak detection algorithm including all the proposed phases is very complex and requires a core with native FPU support because the fixed-point representation decreases the accuracy significantly. Overall, we will show that their results in mJ are 7× higher than our method.

An effective and computationally efficient threshold-based approach for QRS extraction and heart rate (HR) calculation is the Pan and Tompkins algorithm (PT). It relies on an adaptive dual-threshold technique for the R peaks detection [12], leveraging a filtering stage and simple adaptive thresholding methods. PT is a robust technique that uses a pre-processing pipeline that includes standard filtering techniques (pass-band, derivative, squaring, integration). This technique can also be applied to signals with arrhythmia. Furthermore, it can be adapted to execute on real-time streaming data, which is crucial in the context of wearable systems. PT is a standard approach that was proposed several years ago, but

recent works adopt this methodology yet [13,14]. We outperformed the accuracy and energy consumption of these works, optimizing the R peak detection on our target architecture.

In this work, we propose a lightweight design for HR computation based on the PT algorithm. We implement the PT aiming for an acceptable trade-off between computational complexity and energy efficiency. Furthermore, we propose a real-time application for ECG monitoring based on an end-to-end system from the data acquisition to the inference. The proposed methodology is optimized by simulating in Matlab the real-time operation and then implementing it with a multi-board setup. Then, the processing is coded in C language and can work in data-streaming or with an existing dataset. The proposed system provides a power budget of less than 5 mW, for wearable and near sensors processors. We aim to process ECG signal to carry out the HR, which is a crucial physiological parameter to detect anomaly conditions in the heartbeats [15]. Our methodology obtains an acceptable HR detection reliability (higher than 99%) in pathological or sudden changes of the biosignal. The target device that we use for experimental assessment is the Parallel processing Ultra-Low Power (PULP) many-core platform designed for smart ULP embedded devices [16]. For the evaluation, we analyze the performance on the Vega SoC [17], a PULP platform running at 0.8 V at an operating frequency of 170 MHz, and on ARM Cortex-M4, using the STM32 NUCLEO-F401RE development board at 1.8 V and 84 MHz. The PULP provides extreme energy efficiency, and we obtained an energy consumption of 0.2 mJ when considering an average on 25 s of running time. We performed tests on four datasets, three existing ones and one acquired from the proposed system in real-time, taking into account several options: normal

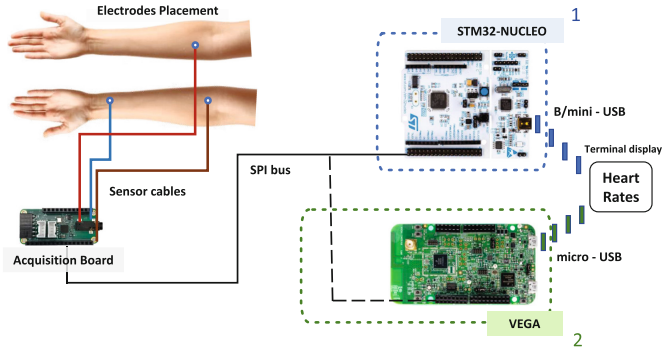


Fig. 1. Hardware diagram of the proposed system. The active electrodes are located on each forearm and one on a wrist, setting Lead I for the collected data. Single-channel ECG is acquired with a custom AFE board (MAX30003), which sends data via SPI to the platform for processing. We consider two alternative designs: (1) STM32NUCLEO for the initial setup and (2) Vega for ULP optimization. Output and communication are managed via a B/mini-USB and a micro-USB cable, respectively, that leads the platform to a terminal to visualize the HR values.

conditions, arrhythmia [18], intense physical exercise [11]. Overall, we achieved accuracy above 99.5% that we compared with other state-of-the-art solutions.

2 Methodology

2.1 System Architecture

This work proposes a modular setup for ECG detection. The acquisition board relies on Maxim MAX30003 [19], a chip for ULP acquisition of ECG. MAX30003 is a complete, biopotential analog front-end solution for wearable applications. It offers high performance for clinical and fitness applications at extreme energy efficiency, reaching 85 μW average power consumption. The analog acquisition is based on a 2 leads differential channel providing ECG waveforms and heart rate detection. The biopotential channel has ESD protection, EMI filtering, internal lead biasing, and DC leads-off detection. The biopotential channel also has high input impedance, low noise, high CMRR, programmable gain, as well as low-pass and high-pass filter options. The digital back end is based on an SPI interface to enable data streaming and communication with an external MCU.

Figure 1 depicts the custom board equipped with MAX30003 and with two alternative test benches: the first one with NUCLEO-F401RE board, used for initial setup and algorithm tuning, and the second one with Vega custom board [17], employed for ULP operation and optimized performance. In both test benches, ECG data are sent from AFE to MCU via SPI. Vega allows a USB device mode interface with a micro USB connector at an operating frequency of 2.4 GHz with a reference oscillator frequency of 32 MHz. Vega receives data from AFE via a 5 MHz SPI channel connection (Vega acts as master). Data loaded via SPI is stored in the Vega L2 memory as 24-bit signed fixed-point numbers, with the

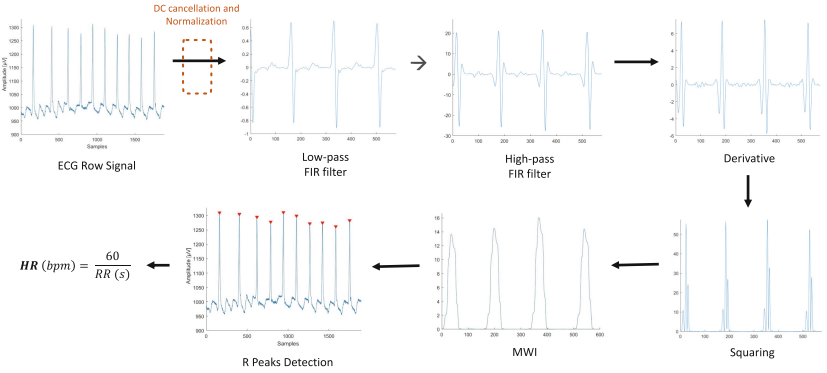


Fig. 2. Signal processing steps based on PT technique: (1) Cancellation of DC component and addition of normalization; (2) Band-pass filter that combines the low- and the high-pass filters; (3) Derivative function; (4) Squaring function; (5) Moving window integrator (MWI); (6) R peaks detection. In the last step we compute the HR in beats per minute.

least significant bit. Acquired ECG samples are used as input of the embedded implementation, described and profiled in Sect. 3.

The presented prototype can be integrated into a single PCB with a 20×10 mm form factor, suitable for minimally obtrusive wearable applications. We use a recent SoC implemented in 22 nm technology, namely Vega [17]. It provides a DSP-oriented instruction set architecture (ISA) based on the Parallel Ultra-Low-Power Platform (PULP) [16]. The PULP SoC is equipped with a 2 MB SRAM scratchpad memory (L2), hosting the resident code and application data. A hardware unit called *μ-DMA* performs autonomous data transfers between the L2 memory and the peripherals. The peripherals and the MCU core reside in different clock domains so that the frequency of each domain can be tuned to sustain the application workload with low power consumption (up to 500 MHz for a 22 nm technology node). The clocks of the peripherals can be further divided to match the operating frequency of slower external devices.

2.2 Algorithm Description

To compute the HR, we adopt a signal processing pipeline based on the PT technique [12]. This methodology adopts a dual-threshold technique to detect the R peaks and includes multiple pre-processing signal steps required to improve the signal analysis. The block diagram is depicted in Fig. 2.

The signal processing pipeline includes a set of pre-processing digital filters followed by the computation of R-peaks. The original implementation considers a sampling rate 200 Hz.

1. **Band-pass filter**. The low-pass component applies a second-order transfer function to the signal obtaining a difference equation with a delay of 5 samples and a DC gain of 36. The high-pass design is characterized by a first-order transfer function, with a delay of 16 points and a gain of 1. Overall, the bandpass filter provides a 3 dB pass-band between about 5 12 Hz and reduces the noise due to the muscle, the baseline wander and the T-wave interference/frequency content. This filter supplies poles and zeros only on the unit circle, so the system is characterized by a minimum phase, a minimum group delay, and better stability. As a final effect, it increases the signal-to-noise ratio.
2. **Derivative**. The signal is differentiated using a 5-points derivative. The result provides information about the slope of the input waveform. This filter introduces a delay of 2 samples and a gain of 0.1.
3. **Squaring**. The output of the derivative signal is squared to enhance the R peaks, leading the signal to the positive y-axis to emphasize the high frequencies that include the R-peaks. This step makes it easier to distinguish R peaks from T-waves.
4. **Integration**. From the output of the squared signal, a moving window integrator extracts the duration of the QRS complex, obtaining a time-averaged signal. Usually, the window length is equivalent to the widest QRS complex (around 150 ms, corresponding to 30 samples 200 Hz). The time of the rising direction of the window is the duration of the QRS complex.

5. **Computation of R peaks**. The final part of the algorithm finds a set of fiducial marks corresponding to the temporal location of the peaks in the integrated signal. Fiducial marks determined in this area are potential candidates for R peaks. An initial phase of the implementation is necessary for the tuning (2 s 128 Hz). The fiducial mark is compared with a threshold value $threshold_{I1}$ that considers the current estimation and both signal and noise peaks:

$$threshold_{I1} = npk_I + 0.25 * (spk_I - npk_I) \tag{1}$$

where npk_I is the estimation for any peak that is not related to an R peak (e.g., the peaks of T waves), and spk_I is the estimated value for the R peak level. When a new peak $peak_I$ is detected, it must be classified as a noise peak or a signal peak. If a sample is greater than the current threshold value $threshold_{I1}$, then it is a peak candidate. In addition, it must have a distance of at least 200 ms from the previously detected peak: this value, referred to as min_rr_width, is the minimum latency time between adjacent R peaks due to physiological constraints. Otherwise, the fiducial mark is considered a noise peak. spk_I and npk_I parameters are updated accordingly:

$$spk_I = 0.125 * peak_I + 0.875 * spk_I \tag{2}$$

$$npk_I = 0.125 * peak_I + 0.875 * npk_I \tag{3}$$

If no R peak candidate is found in an interval of duration $1.66 * max_rr_width$ starting from the previous peak and ending with the current sample, the algorithm performs a search-back operation on this interval the interval using a lower threshold $threshold_{I2}$ that is empirically computed as:

$$threshold_{I2} = 0.5 * threshold_{I1}; \tag{4}$$

The original PT algorithm performs R peaks detection also on the output of the bandpass filter, introducing a set of variables with the same meaning (i.e., $peak_F$, spk_F, npk_F, $threshold_{F1}$, and $threshold_{F2}$). We verified experimentally that this step can be skipped without invalidating the detection quality. If a peak candidate occurs after the 200 ms refractory period but within 360 ms of the previous peak, the algorithm makes an additional check to determine if it is an abnormally prominent T wave. This decision is based on the mean slope of the waveform at that position, which must be greater than one-half that of the previous peak. Finally, the average distance between R peaks is computed as the mean of the eight most-recent RR intervals. The average heart rate can be used to refine the duration of the search back interval.

3 Results

This section provides an experimental evaluation of our system. We use GVSoC [20], an open-source simulator for PULP architectures, to implement and debug the algorithm. GVSoC can simulate a full platform, including multi-memory

levels and multi I/O peripherals, and provides a good trade-off between simulation speed, timing accuracy, and completeness. The average energy consumption for the Vega platform has been derived by a post place-&-route simulation on the RTL. The metrics of interest for our performance analysis are *throughput* (computed as the number of input data samples over the total execution cycles), *energy efficiency* (operations performed in a second over power consumption), *total energy consumption* (in mJ), and *accuracy* of the detection rate (in percentage).

3.1 Implementation on the PULP Platform

The AFE IC, described in Sect. 2.1, is connected to the ECG electrodes using 3 ECG surface sensors: two sensors are placed on the wrist, and the other one around the upper forearm of the subject, as a voltage reference. This setup allows sampling with an 18-bits resolution 128 Hz. In this application, the signal is read 13 samples at a time using a FIFO. We apply the PT algorithm described in Sect. 2.2, implemented in C language, to support different (integer or floating-point) data types. The code supports both buffered and the data-streaming simulation with configurable parameters for the sampling frequency. In the case of buffered execution, the input buffer size is selected to contain at least 1.66 times an R-R interval, considering that the maximum physiological beats per minute are 60 or 80 (max 1.66 beats per second). The code includes buffers for the results of the intermediate filters. These buffers have the exact window size for the corresponding filter and are implemented as circular buffers to reduce memory consumption. Buffered execution can be used to execute the algorithm on pre-recorded ECG datasets, while the streaming variant is more efficient for real-time data acquisition.

In addition to the original PT design, we added a preliminary normalization step that removes the DC drift by subtracting the mean value and then dividing by a maximum absolute value. In the case of buffered execution, this value can be computed as the maximum value in the input buffer; otherwise, we can use the maximum value provided by the sensor as reported in the datasheet. The result is a signal normalized in the range [-1, 1], improving the numerical stability and precision of the next steps.

The filter coefficients are pre-computed using MATLAB and saved into the local memory to maximize the efficiency of the initial steps. To guarantee the minimum latency for streaming execution, we designed a step-by-step convolution function that is invoked for each new available value (i.e., a new input sample or a value computed by the previous filter) and applied a linear convolution filter to the tail on the corresponding data buffer. As introduced in Sect. 2.2, we only consider the integrated signal for R peak detection. Finally, we apply the computation of the HR (beats per minute) from the RR average value. Figure 3 shows an example of the output result of the R peaks detection and the R-R intervals assessment extracted from a segment of the record 232 of the MIT-BIH Arrhythmia Dataset. Even though some fiducial points can be drifted forward

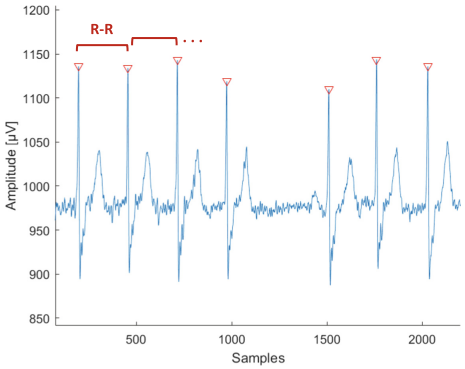


Fig. 3. Output result of the proposed R peaks detection and R-R intervals method from a segment of record 232 characterized by the supraventricular ectopic beats from the MIT-BIH database.

Table 1. Cycles for each sample, instructions, energy efficiency, throughput, and time executing on the target platforms (average values on a 25 s time window).

	Pulp VEGA	Cortex-M4 (processing pipeline)
Cycles	2771	3154
Instructions	2204	3148
Energy efficiency [Gop/s/W]	34.7	3.8
Throughput [samples/ms]	61.35	26.63

or backward by one sample w.r.t. the exact peak positions, this effect does not affect the correct computation of the R-R distance.

3.2 Performance Analysis and Energy Consumption

Table 1 reports the performance parameters executing the program (in streaming mode) on PULP (VEGA SoC) and Cortex-M4 (STM32NUCLEO-F401RE development board) introduced in Sect. 1. We deployed an alternative setup for these experiments where an additional STM32NUCLEO board is used in place of VEGA for the signal processing pipeline. In both cases, the energy consumption of the Nucleo board used for system initialization and debugging is not considered. The resulting values show that execution on the PULP platform is 2.3× faster than Cortex-M4.

Table 2 depicts the energy consumption (in mJ) of our algorithm executed on NUCLEOF401RE and Vega platforms compared to the state-of-the-art solution described by De Giovanni et al. [10], which executes on a PULP platform based on the Mr.Wolf architecture [21]. The operating frequency reported for Cortex-M4 is its maximum frequency. For VEGA, we are using an operating frequency lower than the maximum to make a more fair comparison with state-of-the-art

Table 2. Energy consumption of different SoA solutions for R peak detection (average values on a 25 s time window).

	Platform Architecture	ISA	Algorithm	Technology [nm]	Operating frequency [MHz]	Energy consumption [mJ]
De Giovanni et al. [10]	Pulp (Mr.Wolf)	RV32ICMF + Spec. Ext	Adaptive slope	40	170	1.553
This work	Cortex-M4	ARMv7-M	PT	90	84	2.652
	Pulp (Vega)	RV32ICMX + Spec. Ext	PT	22	170	**0.203**

Table 3. Comparison of R peaks accuracy.

	Acc [%]
Moreira et al. [14]	93.26
De Giovanni et al. [10]	97.90
Tekeste et al. [9]	99.37
Lu et al. [13]	99.41
This work	**99.53**

solutions. The energy consumption has been estimated using an average power consumption reported by the datasheets. Considering an execution time of 25 s, the average energy consumption of our system is almost 7× lower.

3.3 Algorithm Accuracy

Table 3 reports an accuracy comparison between our solution and other works. In the worst case, our algorithm reaches 99.53% on the high-intensity physical exercise dataset [11]. To evaluate the accuracy, we used the MATLAB *findpeaks* function as a golden reference, which returns the local maxima. It is extremely accurate, but it has two main flaws. First, it is computed intensive, which is highly detrimental to its adoption in the ultra-low-power embedded domain. Second, it cannot be adapted to a streaming context, so its adoption would increase the latency of the results. We computed the *accuracy* as follows:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n} (x_{G,i} - x_{PT,i})^2}{n}}; \tag{5}$$

$$Acc = 100 - (\frac{RMSD}{x_{max} - x_{min}}); \tag{6}$$

where x_G and x_{PT} are the RR intervals (in samples) computed using the golden model or the proposed method, respectively. The parameter n is the number of detected RR intervals, and x_{max} and x_{min} are the maximum and minimum in the set of RR interval values.

Figure 4 depicts the accuracy of the code tested on four different datasets. The datasets we consider are Normal Sinus Rhythm (NSR) and Atrial Flutter (AFL). They are both from the MIT-BIH Arrhythmia database, sampled 360 Hz

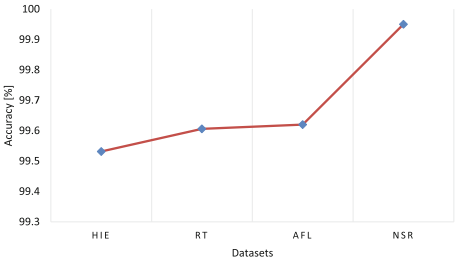


Fig. 4. Accuracy evaluation on four different datasets: High Intensity Exercise (HIE) [11], acquired ECG signal in real-time (RT) with the proposed system design, Atrial Flutter (AFL) [18], and Normal Synus Rhythm (NSR) [18].

[18]. The third is the ECG signal acquired in real-time (RT) from our signal acquisition system (described in Sect. 2.1). Finally, the signal on high-intensity exercise (HIE), sampled 250 Hz [11]. The figure shows the higher value in NSR, for which we achieve 99.95%. In the case of tachyarrhythmia, called atrial flutter (AFL), the accuracy is 99.62%. In the RT, we obtain an accuracy of 99.61%. In HIE, where the beats change suddenly, we assess the lower value of 99.53%.

4 Conclusion

This work presents the design and implementation of a heart-rate detection system leveraging the PT algorithm on low-power MCUs. We consider two alternative platforms, a commercial MCU based on ARM Cortex-M4 and an ultra-low-power solution based on RISC-V, namely the Vega SoC. Experimental results show that our approach is lightweight design and executes in a few thousand cycles. This system provides a lifetime battery of 81 h with a 100 mAh battery, achieving an accuracy comparable to the state-of-the-art solutions and an energy efficiency that is one order of magnitude better. This work does not aim at classifying specific health problems but rather to detect HR in real-time at high reliability and energy efficiency. In future work, we will add machine learning algorithms (e.g., kNN, SVM, or CNN) to the system pipeline with the aim of detecting anomalies in the HR variability, such as arrhythmia or stress conditions. Moreover, we will design a parallel version of the code using the programmable parallel accelerator available on VEGA to further improve performance and energy efficiency compared to commercial alternatives.

Acknowledgements. The work was partially funded by the EU project The European PILOT (g.a. 101034126).

References

1. Rai, P., et al.: Nano- bio- textile sensors with mobile wireless platform for wearable health monitoring of neurological and cardiovascular disorders. J. Electrochem. Soc. **161**(2), 3116–3150 (2013)
2. Lee, S., et al.: Integration of transparent supercapacitors and electrodes using nanostructured metallic glass films for wirelessly rechargeable, skin heat patches. Nano Lett. **20**(7), 4872–4881 (2020)
3. Forooghifar, F., et al.: A self-aware epilepsy monitoring system for realtime epileptic seizure detection. Mob. Netw. Appli. **27**, 1–14 (2019)
4. Benatti, S., et al.: A versatile embedded platform for EMG acquisition and gesture recognition. IEEE Trans. Biomed. Circ. Syst. **9**(5), 620–630 (2015)
5. Magno, M., et al.: Autonomous smartwatch with exible sensors for accurate and continuous mapping of skin temperature. In: 2016 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 337–340 (2016)
6. Schönle, P., et al.: A multi-sensor and parallel processing SoC for miniaturized medical instrumentation. IEEE J. Solid-State Circ. **53**(7), 2076–2087 (2018)
7. Park, J.S., Lee, S.W., Park, U.: R peak detection method using wavelet transform and modified Shannon energy envelope. J. Healthc. Eng. **2017**, 1–14 (2017)
8. Rodrigo, A.M., Alcaraz, R., Rieta, J.: Application of the phasor transform for automatic delineation of single-lead ECG fiducial points. Physiol. Meas. **31**, 1467–1485 (2010)
9. Tekeste, T., et al.: Ultra-low power QRS detection and ECG compression architecture for IoT healthcare devices. IEEE Trans. Circ. Syst. I: Regul. Pap. **66**(2), 669–679 (2019)
10. de Giovanni, E., et al.: Adaptive R-Peak Detection on Wearable ECG Sensors for High-Intensity Exercise (2021)
11. de Giovanni, E., et al.: ECG in High Intensity Exercise Dataset, Nov 2021
12. Pan, J., Tompkins, W.J.: A real-time QRS detection algorithm. IEEE Trans. Biomed. Eng. BME **32**(3), 230–236 (1985)
13. Xuanyu, L., Pan, M., Yang, Yu.: QRS detection based on improved adaptive threshold. J. Healthc. Eng. **2018**, 1–8 (2018)
14. Leite, J., et al.: Online Heartbeat Classification Using Low Cost Algorithms, Mar 2020
15. Kirchhof, P., et al.: 2016 ESC guidelines for the management of atrial fibrillation developed in collaboration with EACTS. Euro. Heart J. **37**(38), 2893–2962 (2016)
16. Rossi, D., et al.: PULP: a parallel ultra low power platform for next generation IoT applications. In: 2015 IEEE Hot Chips 27 Symposium (HCS), pp. 1–39 (2015)
17. Rossi, D., et al.: 4.4 A 1.3TOPS/W @ 32GOPS fully integrated 10- Core SoC for IoT end-nodes with 1.7Î¼WCognitiveWake-Up From MRAMBased state-retentive sleep mode. In: 2021 IEEE International Solid- State Circuits Conference (ISSCC), vol. 64, pp. 60–62 (2021)
18. Moody, G.B., Mark, R.G.: The impact of the MIT-BIH Arrhythmia Database. IEEE Eng. Med. Biol. Mag. **20**(3), 45–50 (2001)
19. Maxim datasheet. https://datasheets.maximintegrated.com/en/ds/MAX30003.pdf., (Accessed 10 May 2022)
20. Bruschi, N., et al.: GVSoC: a highly configurable, fast and accurate full-platform simulator for RISC-V based IoT processors. In: 2021 IEEE 39th International Conference on Computer Design (ICCD), pp. 409–416
21. Pullini, A., et al.: Mr.Wolf: an energy-precision scalable parallel ultra low power SoC for IoT edge processing. IEEE J. Solid-State Circ. **54**(7), 1970–1981 (2019)

Sensors and Sensing Systems

A Non-Hilbertian Inversion Technique for the Diagnosis of Faulty Elements in Antenna Arrays

Valentina Schenone[1], Alessandro Fedeli[1], Claudio Estatico[2], Matteo Pastorino[1], and Andrea Randazzo[1(✉)]

[1] Department of Electrical, Electronic, Telecommunications Engineering and Naval Architecture, University of Genoa, 16145 Genoa, Italy
valentina.schenone@edu.unige.it, {alessandro.fedeli, matteo.pastorino,andrea.randazzo}@unige.it
[2] Department of Mathematics, University of Genoa, 16146 Genoa, Italy
estatico@dima.unige.it

Abstract. Nowadays, antenna arrays are important tools adopted in a great number of applications including radar, mobile and satellite communication systems, and electromagnetic imaging. Moreover, in these applications, arrays with a high number of elements are ever more requested, which results in a growing possibility of damages in the array. The identification of defective components in array of antennas is really significant due to their applicative use: indeed, faulty detected elements can be fixed, thus avoiding to replace the whole antenna. In this work, a diagnostic technique for planar antenna arrays is presented. This approach enables recovering the eventually defective elements of the antenna under test using far-field data. To this end, an inversion approach established outside the standard context of Hilbertian spaces is used to address an inverse-source problem. A numerical validation concerning simple array antennas has been carried out to study the performances of the approach versus some antenna parameters, e.g., the size of the array.

Keywords: Diagnosis of antenna array · Inverse problems · Non-Hilbertian methods

1 Introduction

Arrays of antennas have a wide variety of applications, including telecommunications and microwave imaging [1–5]. Indeed, their main benefit is the capability to fit their radiation properties to the framework in which they operate.

In this scenario, antenna diagnosis is fundamental to identify defects that might result in changes of the radiated fields required by the specific application. Furthermore, current applications often require arrays composed by a high number of antennas. This increases the probability of faulty elements occurrences and makes their identification and replacement a very important task.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 173–181, 2023.
https://doi.org/10.1007/978-3-031-16281-7_17

The antenna array diagnosis may be considered an inverse problem: measured far-field data are used to reconstruct the distribution of currents (or the corresponding feed coefficients) on the inspected antenna [6]. However, this problem requires to face ill-posed equations. Many solutions have been offered by the current scientific research to cope with such an inverse problem. For instance, methods relying upon the so-called equivalent source reconstruction method to reconstruct an equivalent currents distribution through Huygens' principle have been proposed [7–9]. Compressive sensing (CS)-based approaches were also presented [10–14], assuming that failures belong to a sparse distribution.

A new diagnostic approach is devised in this paper to extract the excitations of damaged antennas. This procedure relies on an non-Hilbertian spaces-based technique to perform inversion in a regularized way [15–17]. This family of methods has been able to produce promising reconstruction outcomes with an appropriate tweaking of the exponent function, which allows enforcing sparsity without that sparseness conditions are severely fulfilled [18, 19]. This technique brings some advantages over existing methods. Indeed, as regards procedures based on compressive sensing, they allow obtaining good reconstructions but mainly in the case of sparse solutions. Furthermore, in order to apply these techniques, some specific conditions need to be met, the so-called RIP conditions [10, 18, 19]. On the contrary, inversion in non-Hilbertian spaces frees users from the need to check the RIP conditions. It is also worth mentioning the approaches formulated in the classic Hilbert spaces. The use of these techniques often leads to solutions characterized by significant smoothing effects, which therefore are not suitable for the reconstruction of sparse defects. In this sense, the non-Hilbertian strategy makes it possible to restore different types of solutions, through an appropriate adaptation of the exponent function. The approach is validated through numerical simulations concerning antenna arrays under test of different sizes. More in details, a study has been conducted with two percentage of faulty elements for each size of antenna array.

This contribution is structured as follows. Section 2 presents the involved inverse-source problem and discusses the solving strategy. In Sect. 3, the devised solving methodology is assessed in a numerically simulated framework. At the end, conclusions are given in Sect. 4.

2 Mathematical Formulation

In Fig. 1 the analyzed antenna diagnostic configuration is sketched, where $S = S_x \times S_y$ antennas are centered at points (x_s, y_s). The spacings between antenna elements are δ_x (on the x axis) and δ_y (on the y axis). A working angular frequency ω_0 is considered.

The developed procedure detects eventually present faulty elements from measurements of the radiated electric far-field. In particular, a reference planar array is considered with $\boldsymbol{a}$ being the vector of element feed excitations. As concerns the antenna under test (AUT), it is supposed to have $S_h = hS$ faulty elements in the array (h the percentage of failure) and $\hat{\boldsymbol{a}}$ is the vector of the AUT excitations. Two vectors $\boldsymbol{F}$ and $\widehat{\boldsymbol{F}}$ are obtained for reference array and AUT by measuring the components of the radiated field in D points (θ_d, φ_d), in the far field region. These measurements are collected with an angular interval between measurement points of $\Delta\theta = \Delta\varphi$ and $0 \leq \theta \leq \frac{\pi}{2}$, $0 \leq \varphi < 2\pi$.

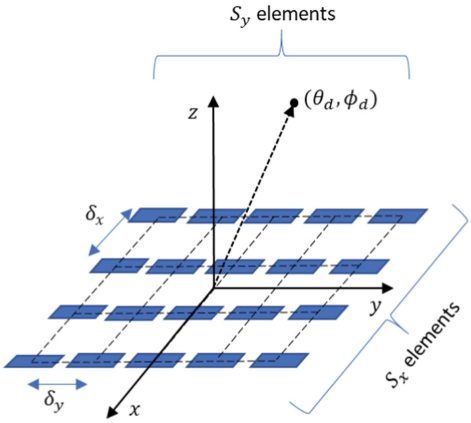


Fig. 1. Sketch of the considered array of antennas.

The radiated far field pattern at the measurement points (θ_d, φ_d), for $d = 1, \ldots, D$, can be written as [20]

$$F^{\boldsymbol{c}}(\theta_d, \varphi_d) = \sum_{s=1}^{S} a_s E^c_{sd}(\theta_d, \varphi_d) e^{j\frac{2\pi}{\lambda}(x_s sin\theta_d cos\varphi_d + y_s sin\theta_d sin\varphi_d)} \tag{1}$$

where $c = \{\theta/\varphi\}$ represents the θ/φ-pattern component and E^c_{sd} represents the embedded pattern for the antenna s in (θ_d, φ_d) [13, 21], which is the array's far-field pattern when the antenna s is stimulated with unitary excitation while other antennas are matched. By combining all the available D observation directions with reference and faulty array measurements, a linear system, which joins the vector $\boldsymbol{H} = \boldsymbol{F} - \widehat{\boldsymbol{F}}$ with the unknown vector of faulty feed coefficients $\boldsymbol{b} = \boldsymbol{a} - \widehat{\boldsymbol{a}}$, is obtained, i.e.,

$$\boldsymbol{H} = [M]\boldsymbol{b} \tag{2}$$

where $[M]$ is an $D \times S$ matrix with elements given by $[M]_{sd} \triangleq E^c_{sd}(\theta_d, \varphi_d) e^{j2\pi(x_s sin\theta_d cos\varphi_d + y_s sin\theta_d sin\varphi_d)/\lambda}$. Therefore, each coefficient of the matrix can be obtained knowing a-priori the array geometry and the embedded pattern measurement for each antenna in the array. It is worth noting that $\boldsymbol{b}$ has zero-value in correspondence of non-faulty elements, whereas non-zero coefficients reveal a damaged antenna and their value provide information about fault entity. Thus, the diagnosis is performed by inverting this equation and recovering $\boldsymbol{b}$.

To this end, a regularization strategy in L^p spaces based on the truncated Landweber algorithm is proposed to solve the inverse problem [22–24]. Specifically, this is an iterative method where, after defining an initial value $\boldsymbol{b}_0$ (e.g., $\boldsymbol{b}_0 = 0$ if no a-priori information is available), at each iteration the solution is updated as follows:

$$\boldsymbol{b}_{k+1} = \mathrm{J}_q\{\mathrm{J}_p(\boldsymbol{b}_k) - \alpha[M]^* \mathrm{J}_p([M]\boldsymbol{b}_k - \boldsymbol{H})\} \tag{3}$$

where $\alpha > 0$ defines the length of each step and is equal to $\alpha = 1/4\left\{1/\|[M]\|_1^2 + (p-1)\left(1/\|[M]\|_2^2 - 1/\|[M]\|_1^2\right)\right\}$, $[M]^*$ is the adjoint of $[M]$, that

is under the examined situations, the Hermitian transposition, and $q = p(p-1)^{-1}$ is the Hölder conjugate of p. The key elements in the truncated Landweber-like method formulated in non-Hilbertian L^p spaces are the so-called duality maps of the L^p space, J_p, used to associate a vector $\boldsymbol{x}$ to its corresponding vector in the dual space, whose l-component is given by $\left\{\mathrm{J}_p(\boldsymbol{x})\right\}_l = |x_l|^{p-1}\mathrm{sign}(x_l)\|\boldsymbol{x}\|_{L^p}^{2-p}$ [22]. In this way, the approach minimizes the residual $R(\boldsymbol{b}_k) = 0.5\|[M]\boldsymbol{b}_k - \boldsymbol{H}\|_p^2$ by moving along non-standard gradient directions in the dual space. The step reported in Eq. (3) is repeated for $k = 0, 1, \ldots, K$ and until a predefined halting criterion is met.

3 Results of Numerical Simulations

A numerical framework has been adopted to validate the devised diagnostic approach. A planar antenna array with one-parameter Taylor excitations is simulated with a sidelobe level equal to $R = 25$ dB [25]. Besides, ideal isotropic antennas are considered, with $\delta_x = \delta_y = \lambda/2$ ($\lambda = 0.15\,m$). $D = 325$ far-field points are considered, with $\Delta\theta = \Delta\varphi = \pi/18$. A zero mean value Gaussian noise with $SNR = 25$ dB is used to corrupt the synthetic data. Moreover, as regards the antenna under test, total failures randomly distributed in the array are considered (i.e., $a_s = 0$ if element s is faulty).

The relative recovery error on the reconstructed vector of faulty feed coefficients, defined as:

$$e_{rec} = \frac{\|\tilde{\boldsymbol{b}} - \boldsymbol{b}^*\|_1}{\|\boldsymbol{b}^*\|_1} \tag{4}$$

where $\tilde{\boldsymbol{b}}$ and $\boldsymbol{b}^*$ stand for the vectors of the recovered and actual faulty feed coefficients, respectively, has been used to assess the correctness of the diagnosis.

Arrays with equal number of elements along the x-axis and y-axis, i.e., $S_x = S_y = N$, are considered, and the behavior of the method versus the parameter p is studied for $1.1 \leq p \leq 2.0$ with numbers of array elements in the range $N \in [15, 20]$. Such a range for the norm parameter has been selected starting from values of p close to 1, which have been demonstrated in imaging application to be useful for sparse solution [15, 16] and choosing as upper bound $p = 2$, which represents the classical method in Hilbert spaces. This study has been accomplished for percentage of failure $h = 0.1$ and $h = 0.3$. As concerns method parameters, the method is stopped when reaches the minimum normalized root mean square error, and $K = 10000$ has been fixed as the upper bound for iteration number.

The values of e_{rec} versus the exponent parameters p achieved by the developed method with $h = 0.1$ are shown in Fig. 2. In this case, the error is minimum for $p = 1.1$ and then increases for higher values of p. Regarding the variation on array dimensions, as can be noticed, by considering the best recovery error, good reconstructions are achieved for all the arrays. In Fig. 3, two cases of the recovered magnitude of the failure vectors of the examined antenna, obtained with the best value of the exponent parameter ($p = 1.1$), are shown. Such figures confirm that the developed inversion strategy enables a good identification of faulty elements. Moreover, the values of the amplitude of the vector of faulty feed coefficients are very close to the actual ones.

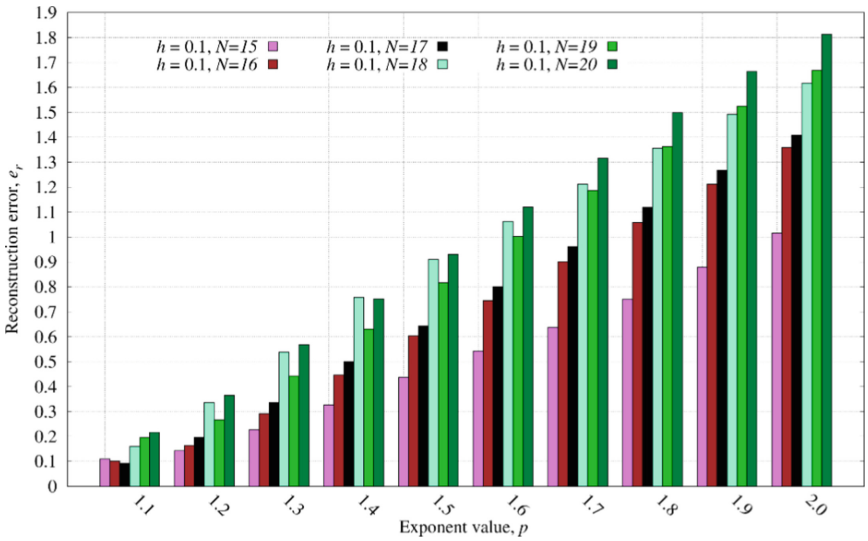


Fig. 2. Relative recovery error for a percentage of failure $h = 0.1$, with array dimension parameter $N \in [15, 20]$ and norm parameter $p \in [1.1, 2]$.

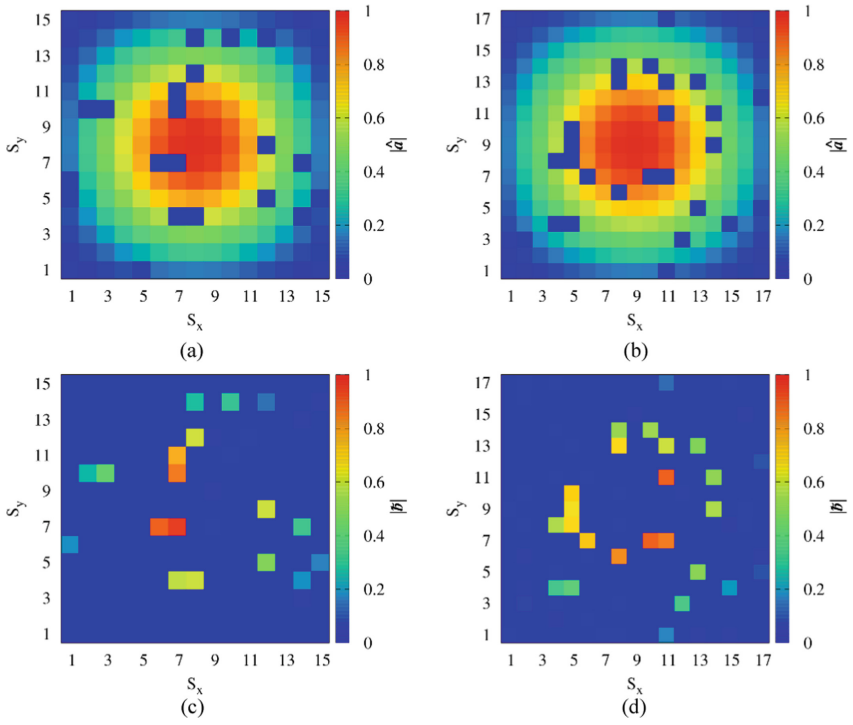


Fig. 3. Feed coefficients magnitude of the damaged antenna for $h = 0.1$ and (a) $N = 15$ and (b) $N = 17$ elements. Best recovered vector of faulty feed coefficients magnitude for (c) $N = 15$ and (d) $N = 17$.

Figure 4 displays the values of e_{rec} against p and the number of array elements with $h = 0.3$. As can be seen, the error has a minimum for a value of $p \in [1.2, 1.3]$ and then it rises again with greater values of p. In this case, by increasing the array dimension, the optimum value of p slightly increases (for $N = 20$, $p = 1.3$). Moreover, a worsening in the reconstruction can be observed as the array dimension increases, since a higher and higher number of non-zero elements has to be reconstructed. Figure 5 shows the magnitude of the actual feed coefficients of the inspected antenna and the best recovered vector of faulty feed coefficients (obtained by setting $p = 1.2$) for $N = 15$ and $N = 17$.

By this analysis, it can be concluded that lower p values enable a good reconstruction when few faulty elements are present (as it happens when $h = 0.1$ and large array are considered), whereas higher values of $p \in [1.2, 1.3]$ perform better for the recovery of unknowns with higher numbers of non-zero elements ($h = 0.3$ and quite large array). This result is consistent with the intrinsic sparsity enhancement of all the methods which minimize the L1 norm of the solution. In our proposal, $p > 1$ exponents close to 1 allow obtaining sparse solutions, while larger values of p are more suited for non-sparse restorations. Therefore, by this study it is evident how this approach allows reconstructing both sparse and non-sparse solutions through an adequate selection of the exponent parameter (differently from the methods relying upon compressive sensing) and without the need of fulfilling the RIP condition. Besides, by comparing the best solution with those achieved with $p = 2.0$, which coincides with the regularization in the classic Hilbert spaces, it can be observed that this technique enables to get better results for all the analyzed cases.

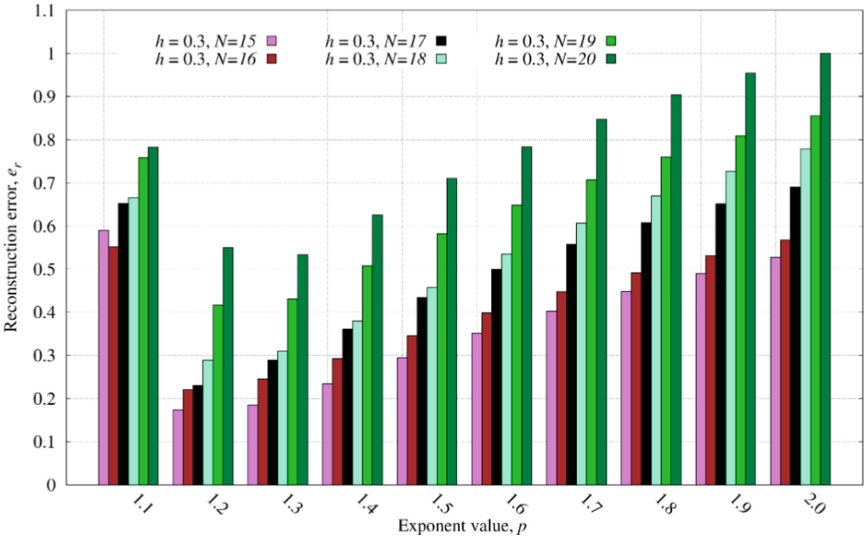


Fig. 4. Relative recovery error for a percentage of failure $h = 0.3$, array dimension parameter $N \in [15, 20]$ and norm parameter $p \in [1.1, 2]$.

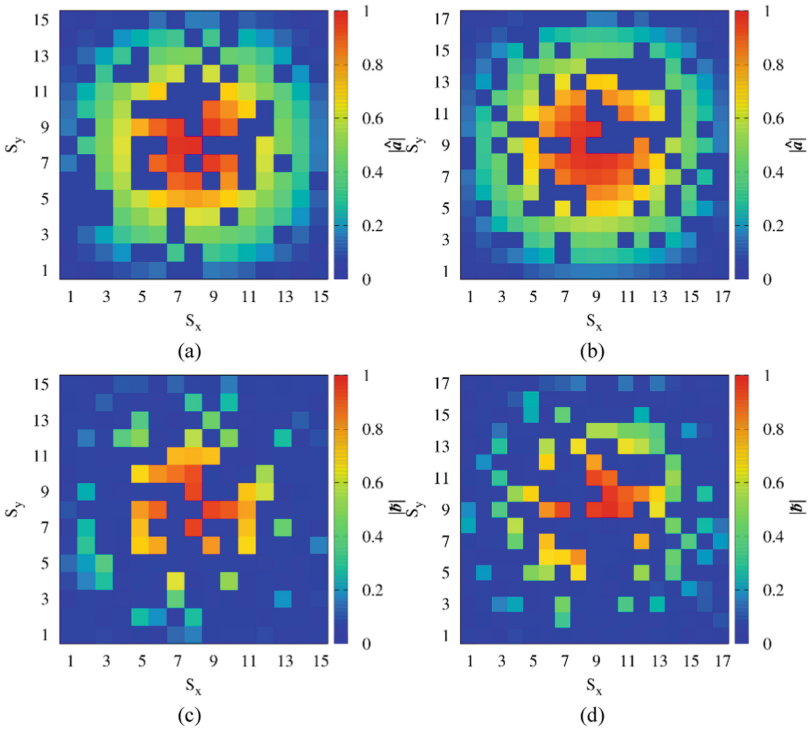


Fig. 5. Feed coefficients magnitude of the damaged antenna for $h = 0.3$ and for (a) $N = 15$ and (b) $N = 17$. Best recovered vector of faulty feed coefficients magnitude (c) $N = 15$, (d) $N = 17$.

4 Conclusion

An antenna array diagnosis method, which aims at solving an inverse-source problem to reconstruct the presence of defective antenna elements in arrays, has been proposed in the present work. In details, a procedure based on the Landweber method in non-Hilbertian L^p spaces was developed to solve the problem at hand in a regularized fashion. Numerical experiments were performed to test the proposed methodology, including a variation of the array size and of the exponent parameter p, considering the test case of a planar antenna array. By fine-tuning the p parameter, accurate diagnostic results can be produced. A comprehensive numerical evaluation of the method and the validation in more realistic settings will be part of future advancements.

Acknowledgements. This work was partially supported by the Italian Ministry for Education, University, and Research under the project PRIN2017 DI-CA, grant number 20177C3WRM.

References

1. Chen, Y., Zhou, W., Yang, S.: Design of a low-profile and low scattering wideband planar phased antenna array. IEEE Trans. Antennas Propag. 69, 8973–8978 (2021)

2. Ahmed, A., Zhang, Y., Burns, D., Huston, D., Xia, T.: Design of UWB antenna for air-coupled impulse ground-penetrating radar. IEEE Geosci. Remote Sens. Lett. **13**, 92–96 (2016)
3. Leone, G., Munno, F., Pierri, R.: Radiation of a circular arc source in a limited angle for nonuniform conformal arrays. IEEE Trans. Antennas Propag. **69**, 4955–4966 (2021)
4. Abbak, M., Nuri Akıncı, M., Ertay, A.O., Özgür, S., Işık, C., Akduman, İ: Wideband compact dipole antenna for microwave imaging applications. IET Microwaves Antennas Propag. **11**, 265–270 (2017)
5. Zeitler, A., Lanteri, J., Pichot, C., Migliaccio, C., Feil, P., Menzel, W.: Folded reflectarrays with shaped beam pattern for foreign object debris detection on runways. IEEE Trans. Antennas Propag. **58**, 3065–3068 (2010)
6. Leone, G., Maisto, M.A., Pierri, R.: Application of inverse source reconstruction to conformal antennas synthesis. IEEE Trans. Antennas Propag. **66**, 1436–1445 (2018)
7. Alvarez, Y., Las-Heras, F., Pino, M.R.: The sources reconstruction method for amplitude-only field measurements. IEEE Trans. Antennas Propag. **58**, 2776–2781 (2010)
8. Konno, K., Asano, S., Umenai, T., Chen, Q.: Diagnosis of array antennas using eigenmode currents and near-field data. IEEE Trans. Antennas Propag. **66**, 5982–5989 (2018)
9. Tzoulis, A., Eibert, T.F.: A hybrid FEBI-MLFMM-UTD method for numerical solutions of electromagnetic problems including arbitrarily shaped and electrically large objects. IEEE Trans. Antennas Propag. **53**, 3358–3366 (2005)
10. Fuchs, B., Coq, L.L., Migliore, M.D.: Fast antenna array diagnosis from a small number of far-field measurements. IEEE Trans. Antennas Propag. **64**, 2227–2235 (2016)
11. Xiong, C., Xiao, G., Hou, Y., Hameed, M.: A compressed sensing-based element failure diagnosis method for phased array antenna during beam steering. IEEE Antennas Wirel. Propag. Lett. **18**, 1756–1760 (2019)
12. Ince, T., Ögücü, G.: Array failure diagnosis using nonconvex compressed sensing. IEEE Antennas Wirel. Propag. Lett. **15**, 992–995 (2016)
13. Salucci, M., Gelmini, A., Oliveri, G., Massa, A.: Planar array diagnosis by means of an advanced Bayesian compressive processing. IEEE Trans. Antennas Propag. **66**, 5892–5906 (2018)
14. Palmeri, R., Isernia, T., Morabito, A.F.: Diagnosis of planar arrays through phaseless measurements and sparsity promotion. IEEE Antennas Wirel. Propag. Lett. **18**, 1273–1277 (2019)
15. Estatico, C., Fedeli, A., Pastorino, M., Randazzo, A.: Microwave imaging of elliptically shaped dielectric cylinders by means of an Lp Banach-space inversion algorithm. Meas. Sci. Technol. **24**, 074017 (2013)
16. Estatico, C., Fedeli, A., Pastorino, M., Randazzo, A., Tavanti, E.: A phaseless microwave imaging approach based on a Lebesgue-space inversion algorithm. IEEE Trans. Antennas Propag. **68**, 8091–8103 (2020)
17. Fedeli, A., Schenone, V., Randazzo, A., Pastorino, M., Henriksson, T., Semenov, S.: Nonlinear S-parameters inversion for stroke imaging. IEEE Trans. Microw. Theory Tech. **69**, 1760–1771 (2021)
18. Massa, A., Rocca, P., Oliveri, G.: Compressive Sensing in Electromagnetics - A Review. IEEE Antennas Propag. Mag. **57**, 224–238 (2015)
19. Tipping, M.E.: Sparse bayesian learning and the relevance vector machine. J. Mach. Learn. Res. **1**, 211–244 (2001)
20. Balanis, C.A.: Antenna theory: Analysis and design. John Wiley & Sons, Hoboken, NJ (2016)
21. Kelley, D.F., Stutzman, W.L.: Array antenna pattern modeling methods that include mutual coupling effects. IEEE Trans. Antennas Propag. **41**, 1625–1632 (1993)
22. Estatico, C., Pastorino, M., Randazzo, A.: A novel microwave imaging approach based on regularization in Lp Banach spaces. IEEE Trans. Antennas Propag. **60**, 3373–3381 (2012)

23. Brianzi, P., Di Benedetto, F., Estatico, C.: Improvement of space-invariant image deblurring by preconditioned Landweber iterations. SIAM J. Sci. Comput. **30**, 1430–1458 (2008)
24. Fedeli, A., Pastorino, M., Ponti, C., Randazzo, A., Schettini, G.: Through-the-Wall Microwave Imaging: Forward and Inverse Scattering Modeling. Sensors. **20**, 2865 (2020)
25. Orfanidis, S.J.: Electromagnetic Waves and Antennas (2016). http://www.ece.rutgers.edu/~orfanidi/ewa/

A Passive, Wireless Sensor Node for Material-Integrated Strain and Temperature Measurements in Glass Fiber Reinforced Composites

Lukas Bertram[1(✉)], Michael Brink[2,3], Klaus-Dieter Thoben[2,3], and Walter Lang[1]

[1] Institute for Microsystems, -Sensors and -Actuators, University of Bremen, 28359 Bremen, Germany
lbertram@imsas.uni-bremen.de
[2] Bremer Institut für Produktion und Logistik GmbH, 28359 Bremen, Germany
[3] Institute for Integrated Product Development, University of Bremen, 28359 Bremen, Germany

Abstract. During production of fiber reinforced plastic (FRP) components, precise measurement and regulation of temperature is important for correct and predictable fabrication. Especially for thicker layer build-ups, material-integrated temperature measurement can help to optimize the fabrication process. During component usage, material-integrated measurement of mechanical forces can be used to monitor component stress and possibly predict failure, in context of condition or structural health monitoring (SHM). Integration of wired sensors can be problematic e.g. due to matrix weakening and wire sensitivity. This paper presents a novel approach for a wireless, material-integrated, batteryless sensor system for both temperature and strain measurement, during both the production process and component application. It uses a flexible printed circuit board, a full bridge strain gauge configuration, a digital temperature sensor and off-the-shelf RFID technology for data and energy transfer, and it is designed to simplify automatic placement during automated FRP production. A case study of bending and temperature tests proves system concept and shows feasibility of the presented approach, though effects of material integration on the system and host component have yet to be analyzed.

Keywords: Wireless sensor · RFID · Material integration · Strain gauge · Temperature · Measurement · Composites · FRP · Structural health monitoring

1 Introduction

Fiber reinforced plastic (FRP) components are increasingly used in various applications due to unique possibilities regarding synergetic combination of material properties, such as light weight and strength. This is made possible by combining

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 182–193, 2023.
https://doi.org/10.1007/978-3-031-16281-7_18

highly rigid and strong fibers, which are able to bear high loads, and polymer matrix material, which fixes the fibers in position and protects them against exterior influences [16]. Production of FRP components is a complex process and resulting material parameters are greatly influenced by various parameters, such as fiber orientation, matrix resin composition, flow rates, curing temperature and many others [11].

1.1 Curing Temperature

One important process parameter is matrix temperature during curing, which describes the process of hardening the fluid matrix material after it has been put into a corresponding mold and been combined with the fibers arranged in the mold. Accurate monitoring and regulation of curing temperature is important for prediction of material parameters and process optimization [11], as it greatly affects the speed and quality of the curing reaction. Especially inside thicker components made of many fiber layers, temperature can vary due to non-uniform heating, heat capacity and thermal conductivity of the matrix material, and the exothermic nature of the resin curing reaction [11]. Material-integrated temperature measurement therefore can be a way to better monitor the curing process.

1.2 Mechanical Loading

Just as structural components made from other materials, FRP components are subjected to different mechanical forces and load scenarios during application, creating stress and wear symptoms over a component's respective life cycle. One of the possible physical quantities usable for component load measurement is mechanical strain (ϵ), which is defined by [5] as ratio of load induced change in length Δl to the original length l. A common way to quantify strain is measuring deformation-induced changes in electrical resistance of a conductor or semiconductor, which are mainly caused by the piezoresistive effect [12]. Strain gauges are a common implementation of this measurement principle. As numerous publications have shown, strain sensors of different designs can be used to monitor the structural health of a component over time, consequently allowing for predictive maintenance. The following paragraph will give some examples:

1.3 State of the Art

Hongki et al. created a high sensitivity strain sensor for structural health monitoring (SHM) of steel bridges, utilizing strain gauges with a Wheatstone bridge circuit [10]. By harvesting power from ambient sources, i.e. solar radiation, the system was able to autonomously generate measurement data and wirelessly transmit these to a base station. Similarly, [3] devised a strain gauge based system for SHM of buildings. Apart from using strain gauges, several publications have used sensoric antennas for strain measurement, as reviewed in [20]. Even though sensoric antenna tag designs have their respective advantages, i.e. very low production cost due to simple design, the most severe disadvantage of this

approach is its large susceptibility to environmental influences, which directly influence the measured entity itself [20]. This can complicate differentiation of disturbances and measurement signal.

Another widely researched approach for monitoring of strain and temperature, especially in context of FRP, uses optical fibers, as [21] and [17] comprehensively review. Though these systems are able to provide in-depth knowledge about several component condition parameters and are suitable for material integration, they require complex analysis equipment and are comparably costly [17]. Concerning material integration, both temperature and strain sensors of different designs have been integrated into FRP materials with good measurement results. Both [4] and [7] proved that structural integration of wired strain gauges into FRP components is a viable method for strain measurement. As of now, these systems have been using wired sensor elements, creating problems and additional effort for integration, such as necessary steps to ensure air tightness at wire positions during resin infusion, weakening of structural integrity due to wire presence, and sensitivity of the sensor system to wire breakage. Broken wires usually cannot be replaced, rendering the whole sensor system useless.

To overcome these limitations, wireless sensor nodes can be integrated into FRP, though not much work about this has yet been published. [18] showed that wireless power transfer via inductive coupling can be used for SHM applications, while [19] proved that integration of passive wireless sensors into FRP is a feasible approach in general. They integrated several wireless sensors that were able to measure temperature and local resin pressure, all aiming at process monitoring, using radio frequency identification (RFID) technology. One approach was able to measure temperature and local resin pressure, but was only functional during production, as most of the electronics needed to be removed after curing. Two other systems, one for temperature and one for temperature and pressure measurement, were wholly integrated into the laminate structure, remaining inside and functional during and after curing. According to the data sheet of the corresponding micro controller, the system is limited to operating temperatures up to 70 °C [8]. For process monitoring, [6] have presented an entirely different approach, where received signal strength of a passive 865 MHz RFID transponder is used as measurement signal. In a vacuum assisted infusion process, they were able to track the flow front of the resin through a glass fiber lay-up, but signal changes during cure were not sufficiently significant to indicate cure status.

1.4 Conducted Work

This paper presents a novel approach for a passive sensor system, for both temperature and strain monitoring, which wirelessly transmits measurement data from inside the component material. The system synergetically combines several capabilities, thereby allowing for usage in process monitoring during FRP production, as well as for condition monitoring during component application. As temperatures during curing can be high, depending on the matrix components used [7], the system is designed for working temperatures of at least 100 °C. To increase system lifetime and minimize system size, the system is powered only by an electromagnetic RFID field, eliminating the need for a battery. The system

also aims to minimize the foreign body effect on the host component by reducing size and incorporating structural holes to allow formation of resin bridges throughout the sensor substrate, similar to the ideas for reduction of the 'wound effect' presented by [13]. Additionally, system design facilitates automatic application during FRP production by aiming for a compact geometry similar to RFID tags widely used in other technology sectors. Thereby it enables usage in larger scale FRP production with higher degrees of automation as envisioned by [1]. In order to reduce system cost and increase compatibility, the sensor system presented uses only commercially available components and standardized RFID communication standards (ISO 14443A, 13.56 MHz).[1]

The case study presented at the end of this paper shows promising results for feasibility of the system concept. Integration tests prove that the system is able to withstand environmental conditions present during vacuum infusion process and curing temperatures (about 80 °C), while remaining functional throughout the whole process. Bending tests of a system prototype glued to a polystyrene specimen show a good correlation between applied force and sensor response. Temperature experiments prove system functionality up to a temperature of 125 °C with a temperature measurement error of less than 1 °C up to 110 °C compared to reference measurements.

2 System Concept

This section gives an overview of the system concept and design guidelines. Description of electronic components and fabrication methods is given in Sect. 3.

Apart from capability to measure strain and temperature, system design was guided by several further considerations, due to the intended application, i.e. usage in industrially produced FRP components. As temperature monitoring is especially feasible for thicker FRP layer configurations, a minimum reading distance of 20 mm is targeted. In order to withstand temperatures during curing, a wide working temperature range was basis for all decisions regarding system concept and component selection. The latter was conducted following the idea of wide compatibility and availability, by choosing standardized electronic components and established RFID technology, allowing easy communication with off-the-shelf RFID readers. Expandability of the system platform for future incorporation of additional functionalities, i.e. different sensors was another aspect during system design.

A schematic overview of the sensor system is given in Fig. 1. Central part of the tag circuitry is an *NHS3152* by NXP, which, in a single package, comprises both the micro controller with the CPU and peripheral interfaces and the RFID interface. The micro controller communicates with the other integrated circuits on the tag and manages measurements, while the RFID interface communicates with the micro controller and the external reader. Also, it harvests energy from the

[1] Selection of the RFID frequency was based on the amount of transferable power, as the transponder is designed to work on harvested power only.

RFID field in order to supply the CPU and the external measurement circuitry for strain and temperature, providing supply voltage from a 'high-drive' GPIO pin.[2]

2.1 Temperature Measurement

Measurement of temperature is conducted via a digital temperature sensor from Texas Instruments (*TMP117*). It is powered via the *NHS3152* (see Sect. 2) and communicates its measurement data via an I^2C connection. It was chosen for its high measurement precision, small size and very low minimum supply voltage of 1.8 V. Compared with alternative methods, i.e. usage of an NTC or Pt100 resistor, the main advantage of this approach is unnecessity of user calibration, as the sensor is factory calibrated.

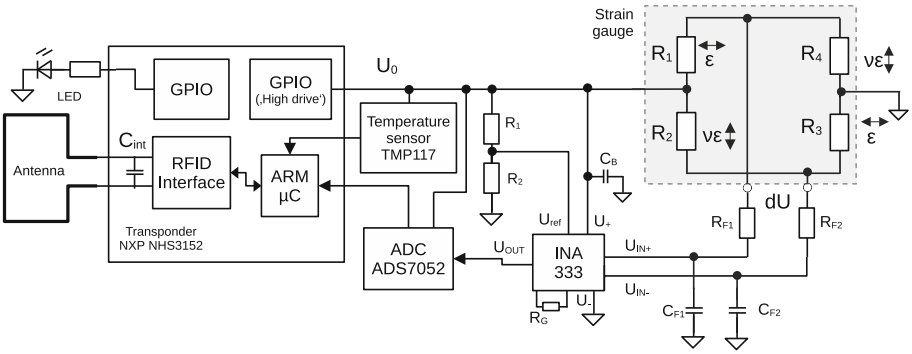


Fig. 1. Simplified block diagram of sensor tag circuit

2.2 Strain Measurement

For strain (ϵ) measurement, a full bridge strain gauge configuration with four active gauges was selected. This configuration combines several advantages over other configurations, the most prominent being intrinsic compensation of temperature influence on gauge resistance and high output voltage. In order to minimize power consumption, a high bridge resistance was targeted. The four gauges of the bridge are laid out to allow bidirectional strain measurement, by orienting two parallel gauges for one direction perpendicular to the two other parallel gauges for the other direction (see Fig. 1). Straining the bridge in one direction increases bridge output, while strain perpendicular to that direction decreases bridge output. Inherent in this configuration is superimposition of lateral contraction ($\nu\epsilon$, occurring perpendicular to component strain). Depending on material properties, this effect influences measurement signal, as contraction perpendicular to strain in main loading direction also affects bridge output in the same way, and therefore needs to be taken into account during measurement data processing.

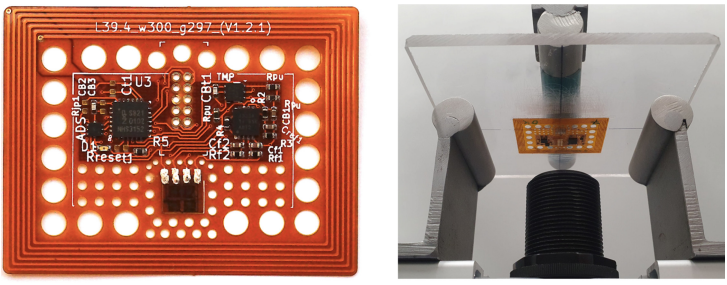
[2] General purpose input output (GPIO). As stated in the data sheet [2], these pins are able to provide currents of up to 20 mA, depending on energy source.

Bridge voltage is first low pass filtered and then amplified by an instrumentation amplifier (*INA333*) with fixed gain. Gain can be selected via R_G. For digitization, it is offset by half the supply voltage via the U_{Ref} pin. This offset is needed as the amplified bridge voltage is measured by an analog to digital converter (ADC) in single ended configuration. It is generated via a high resistance, low tolerance voltage divider (R_1/R_2). A 14 Bit ADC (*ADS7052*) is used, which samples input voltage between ground and supply voltage and communicates measurement data to the micro controller via an SPI connection. As the ADC has a high sampling rate, several measurements can be taken in quick succession, which can be averaged by the micro controller to increase accuracy.

3 Implementation and Fabrication

Tag substrate is a flexible, double layer printed circuit board (PCB) made from layers of polyimide and copper. Figure 2a shows a prototype of the current design. It has a rectangular shape of 29.6 mm × 39.4 mm and a total thickness of about 200 μm (PCB only). In contrast to a circular shape, the shape chosen allows for easy optical identification of tag orientation, which is directly connected to direction of strain measurement. To reduce potential of fracture generation at sharp corners, the rectangle corners have been rounded off.

The electronic components of the sensor are located in the rectangle center, enclosed by the RFID antenna coil, which is placed around the tag circumference. For all electronic components, small package sizes were chosen to minimize the foreign body effect generated. Electronics and the hatched ground pour, shielding them from underneath, were distanced from the antenna coil by a margin of 4 mm, in order to keep antenna Q-factor high (as recommended by [9]). The resulting unoccupied PCB space is covered with holes for better integration into the host component matrix, by allowing fluid resin to pass through during infusion. For assembly, the electronic components were placed onto the PCB and fixed with soldering paste, which was then soldered by heating the tag on a hotplate. At its thickest point, heights of the micro controller and PCB add up to a maximum total thickness of 1.2 mm.



(a) (b)

Fig. 2. a) Flexible PCB sensor tag prototype, fully assembled. b) Test specimen in bending machine

3.1 Strain Sensor

The strain gauge selected is a miniature 5 kΩ full bridge [15], comprising 4 Karma-alloy strain gauges on a 4 mm× 4 mm polyimide thin film, covered by a protective epoxy encapsulation. It provides a gauge factor of 2.1 and allows for ±1.8‰ strain for cyclic loading. To ensure secure installation on the PCB and optimal strain transfer, the gauge is applied with special adhesive (*M-Bond AE10* by Vishay Precision Group). The gauge is positioned symmetrically on the y-axis of the tag rectangle, with as much distance from the electrical components and from the antenna as possible, in order to minimize influence on strain measurement. For the same reason, no components were placed on the PCB along the measurement axes. To optimize mechanical coupling with the surrounding matrix material, holes were placed in the tag substrate around the strain gauges. For electrical contacting, conductive glue was used *(Elecolit 414* by Panacol GmbH). As opposed to soldered connections, the glue chosen remains flexible after curing, thereby minimizing influence of the connection on strain measurement.

4 Results

The following section presents results of experiments conducted with sensor tag prototypes, in order to prove system functionality. Initial integration tests have also been performed, proving the system can withstand conditions inside a vacuum assisted infusion setup, but in-depth analysis of material-integrated measurement data is still subject of ongoing research.

4.1 Antenna and Overall System

During design, a resonance frequency of 13.56 MHz was targeted according to the corresponding RFID standard (*ISO 14443A*). For validation, the resulting resonance frequency was measured using an impedance analyzer and a loop antenna, as described in [2]. The finished tag showed a resonance frequency of 13.642 MHz. For supply of the measurement circuitry via the GPIO pin (see Sect. 2), an average output voltage of 1.7–1.8 V could be measured, providing power of about 0.8 mW. For all experiments, an RFID reader from MicroSensys GmbH was used (*M30 RFID Head*), yielding a maximum reading distance of 43 mm for the fully assembled tag (in free air, 21 °C).

4.2 Temperature Tests

In order to prove temperature measurement functionality and determine maximum operating temperature, the sensor tag was subjected to temperatures from 20–130 °C. For this, it was placed onto a glass surface, alongside a reference sensor, inside a closed glass housing to minimize exterior influences due to air circulation. The glass housing was placed onto a hotplate, while the RFID reader

was mounted above, at distance of 20 mm to the tag. For reference measurement, a Pt100 resistor was connected to a precision multimeter in 4-wire-configuration and attached close to the tag's temperature sensor. Comparing reference temperature with and without RFID field presence, no influence of the reader field on the reference measurement could be found.

The temperature experiment was conducted in two stages. First, hotplate temperature was set to 90 °C, to achieve a temperature similar to a curing process. After 15 min (900 s), hotplate temperature was set to 120 °C, and after 35 min temperature was increased again to 150 °C in order to measure dropout temperature. Results of the experiment are shown in Fig. 3a–3d. From 20–110 °C, absolute temperature error compared to reference is smaller than 1 °C, thereby showing good linearity in the respective range. Above 110 °C, temperature error increases to a maximum of 2.04 °C at 124.7 °C. At a reference temperature of 125.9 °C the first sensor tag dropout was recorded, as data could not be read, marking maximum operating temperature for this experiment.

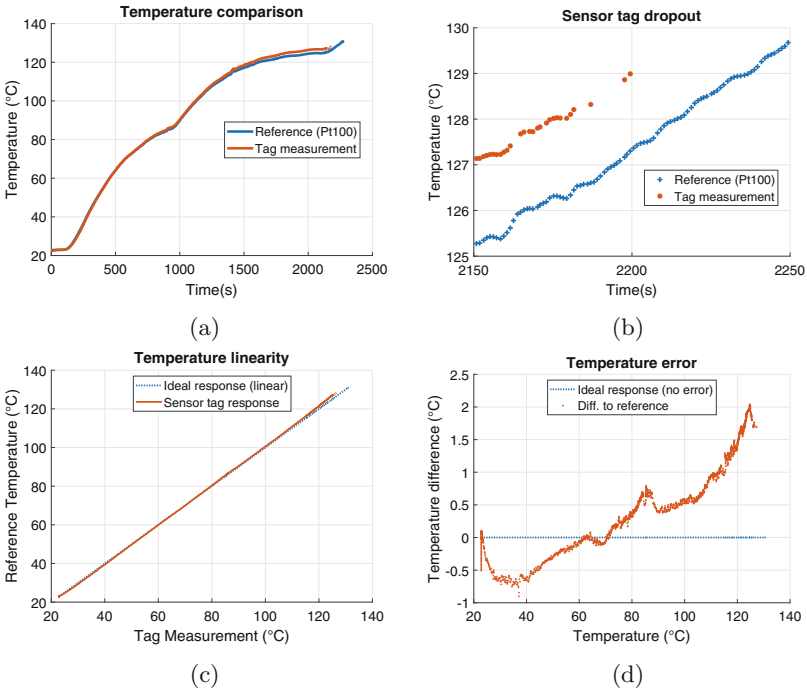
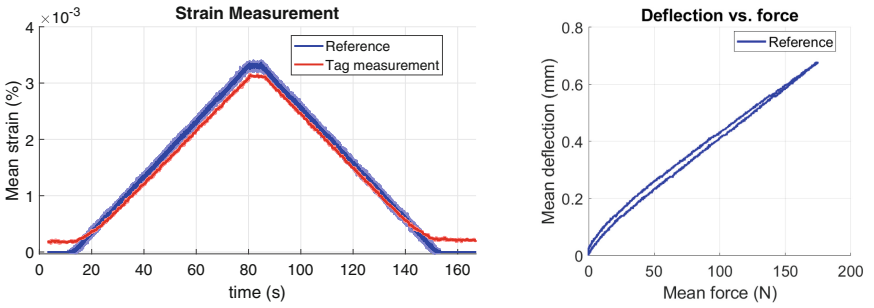


Fig. 3. a) Temperature measurements over time. b) Zoom to sensor tag dropout region (first missing data at 125.9 °C). c) Sensor tag response compared to ideal response. d) Measurement error vs. reference temperature.

4.3 Bending Tests

For validation of strain measurement functionality, a fully assembled sensor tag was mounted onto a 5mm strong bending sample of transparent polystyrene, using cyanacrylate glue. This sample was bent in a 3-point-bending test configuration (see Fig. 2b), using a maximum force of 175 N and a deflection speed of 0.1 mm/min. Bending was performed in both axes, i.e. X and Y, while tag data and data from the bending machine were simultaneously recorded.

Results for strain in X direction are shown in Fig. 4a and 4b. The solid lines depict the averaged strain measurements for three force cycles, while the shaded areas mark the resulting standard deviation at this point. For both bending directions, good correlation of tag and reference measurements can be observed, though difference grows with rising force. Also, tag measurement shows an offset without any strain actually applied to the specimen. For bending in X direction, maximum deviation of tag measurement from reference was –0.24‰ during bending, while the maximum error for Y was 0.7‰. Before and after bending, strain measurement showed an offset, deviating from reference for about 1.8‰ before and 0.25‰ after bending. This difference in offset, along with the differences in deviation for loading and relaxation, may be explained with a material-inherent hysteresical behavior. Figure 4b illustrates this by correlating deflection and force, as measured by the bending machine. It can be seen that the force needed for a certain deflection differs for loading and relaxation, which could be explained by a temporary plastic deformation of the specimen material. Also it is likely that strain may not perfectly be transferred from specimen to tag and from tag to strain gauge, due to flexibility of the glued connections. The general offset of tag measurement may be caused by residual deformation applied to the strain gauge and the tag itself during gluing processes.



(a) (b)

Fig. 4. Mean results and standard deviation (shaded areas) of bending test for n = 3 cycles, F = 175 N, strained in X direction. a) Strain measured by tag and bending machine vs. time. b) deflection vs. force, measured by bending machine.

5 Conclusions

A wireless, passive sensor system for material-integrated measurement of strain and temperature was successfully developed and built. A maximum difference of 0.25‰ between tag and reference strain data for bending in X-direction could be measured in bending tests using polystyrene specimen. Bending in Y-direction yielded a maximum error of 0.7‰. Temperature tests in free air prove correct tag function up to 125.9 °C, while error of temperature measurement was less than 1 °C from 20–110 °C and less than 2.05 °C from 110–125.9 °C. In the future, an in-depth characterization of sensor response, both for strain and temperature is necessary to quantify its accuracy and environmental limitations. Initial integration tests verified applicability of the system for integration into glass fiber reinforced material and measurement of process parameters, i.e. temperature, during all production stages. Results of these experiments will be subject of a future publication.

Characterization of the tags integrated into FRP is yet to be conducted in order to assess influence of surrounding material and the integration process. Also, effects of sensor presence on the surrounding material and structural integrity of the host component need to be studied, as the sensor likely disturbs fiber alignment.

As [14] have shown, presence of carbon fibers in vicinity of RFID antennas degrade antenna performance and power transmission. The system presented in this paper targets integration into the laminate structure, so it is not applicable for usage in carbon fiber based FRP. Opposed to that, non-conducting fiber materials (e.g. aramid or natural fibers) would not inhibit energy transfer, and therefore could be used in combination with the presented system.

Acknowledgements. This work was supported by the German *Federal Ministry for Economic Affairs and Climate Action* based on a resolution of the German Bundestag (Vorh. Nr. 20941 N).

References

1. Brink, M., Berkemeyer, F., Ohlendorf, J.H., Dumstorff, G., Thoben, K.D., Lang, W.: Challenges and opportunities of RFID sensortags integration by fibre-reinforced plastic components production. Procedia Manufacturing **24**, 54–59 (2018). https://doi.org/10.1016/j.promfg.2018.06.008. https://linkinghub.elsevier.com/retrieve/pii/S2351978918305250
2. B.V., N.S.: NTAG Antenna Design Guide (2018). https://community.nxp.com/pwmxy87654/attachments/pwmxy87654/nfc/6155/1/AN11276.pdf
3. Choi, H., Choi, S., Cha, H.: Structural Health Monitoring system based on strain gauge enabled wireless sensor nodes. In: 2008 5th International Conference on Networked Sensing Systems, pp. 211–214 (2008). https://doi.org/10.1109/INSS.2008.4610888

4. Dumstorff, G.: Modellierung und experimentelle Untersuchung von materialintegrierten Sensoren. Ph.D. thesis, University of Bremen, Bremen (2015). https://nbn-resolving.de/urn:nbn:de:gbv:46--00105038-17
5. Grellmann, W., Seidler, S., Altstädt, V., (eds.): Kunststoffprüfung, 3, [aktualisierte], aufl Hanser, München (2015)
6. Hardi, E., Veigt, M., Koerdt, M., Herrmann, A.S., Freitag, M.: Monitoring of the vacuum infusion process by integrated RFID transponder. Proc. Manuf. **52**, 20–25 (2020). https://doi.org/10.1016/j.promfg.2020.11.005. URL https://linkinghub.elsevier.com/retrieve/pii/S2351978920321466
7. Horoschenkoff, A. (ed.): Structural integration of strain gages. Hottinger Baldwin Messtechnik GmbH (2006). http://2007.parcfd.org/seminar/strain-gage/integ.pdf
8. Inc. TI: RF430FRL15xH NFCISO15693 Sensor Transponder (2014). https://www.ti.com/lit/ds/symlink/rf430frl152h.pdf
9. Inc. TI: TI - Passive NFC Temperature Patch Reference Design (2016). https://www.ti.com/lit/ug/tidubt8/tidubt8.pdf?ts=1649853303540
10. Jo, H., Park, J.W., Jeong, H.J.: Develoment of high-sensitivity wireless strain sensor for structural health monitoring. Smart Struct. Syst. **11**(5), 477–496 (2013). https://doi.org/10.12989/SSS.2013.11.5.477, http://doi.org/10.12989/SSS.2013.11.5.477
11. Konstantopoulos, S., Fauster, E., Schledjewski, R.: Monitoring the production of FRP composites: a review of in-line sensing methods. Expr. Polym. Lett. **8**(11), 823–840 (2014). https://doi.org/10.3144/expresspolymlett.2014.84, http://www.expresspolymlett.com/letolt.php?file=EPL-0005435&mi=c
12. Lang, W.: Sensors and Measurement Systems. River Publishers Series in Electronic Materials and Devices. River Publishers, Aalborg (2019). https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=2018166&site=ehost-live
13. Lang, W., et al.: From embedded sensors to sensorial materials—The road to function scale integration. Sens. Actuators A: Phys. **171**(1), 3–11 (2011). https://doi.org/10.1016/j.sna.2011.03.061, https://linkinghub.elsevier.com/retrieve/pii/S0924424711002317
14. Mayordomo, I., Dräger, T., Bernhard, J.: Technical challenges for the integration of passive HF RFID technology in FRP composite materials. In: 2011 IEEE International Conference on RFID-Technologies and Applications, pp. 445–450 (2011). https://doi.org/10.1109/RFID-TA.2011.6068676
15. MicroMesurements, V.P.G.: Datasheet S5020 - Transducer-Class®Strain Gages with Advanced Sensors Technology (2019). https://docs.micro-measurements.com/?id=2622
16. Neitzel, M., Mitschang, P., Breuer, U., (eds.): Handbuch Verbundwerkstoffe: Werkstoffe, Verarbeitung, Anwendung, 2, aktualisierte und erw, aufl Hanser, München (2014)
17. Rocha, H., Semprimoschnig, C., Nunes, J.: Sensors for process and structural health monitoring of aerospace composites: a review. Eng. Struct. **237**, 112231 (2021). https://doi.org/10.1016/j.engstruct.2021.112231
18. Salas, M., Focke, O., Herrmann, A.S., Lang, W.: Wireless power transmission for structural health monitoring of fiber-reinforced-composite materials. IEEE Sens. J. **14**(7), 2171–2176 (2014). https://doi.org/10.1109/JSEN.2014.2305578, http://ieeexplore.ieee.org/document/6736058/
19. Vargas Gleason, M.G., Jedermann, R., Dimassi, A., Lang, W.: Embedded wireless sensor systems for resin flow monitoring in glass and carbon fiber composites. IEEE Sens. J. **19**(22), 10654–10661 (2019). https://doi.org/10.1109/JSEN.2019.2928635

20. Zhang, J., Tian, G., Marindra, A., Sunny, A., Zhao, A.: A review of passive RFID tag antenna-based sensors and systems for structural health monitoring applications. Sensors **17**(2), 265 (2017). https://doi.org/10.3390/s17020265, http://www.mdpi.com/1424-8220/17/2/265
21. Zhou, G., Sim, L.M.: Damage detection and assessment in fibre-reinforced composite structures with embedded fibre optic sensors-review. Smart Mater. Struct. **11**(6), 925–939 (2002). https://doi.org/10.1088/0964-1726/11/6/314, https://iopscience.iop.org/article/10.1088/0964-1726/11/6/314

Multi-camera Metrology System for Shape and Position Correction of Large Fuselage Components in Aircraft Assembly

Daniel Valencia(✉), Jörg Wollnack, Sanjay Kamath, and Leander Brieskorn

Fraunhofer Institute for Manufacturing Technology and Advanced Materials IFAM, 21684 Stade, Germany
daniel.valencia@ifam.fraunhofer.de

Abstract. The assembly of large-volume fuselage components made of composite materials are particularly suitable for metrology-based full automation. So far, mobile coordinate measurement machines have performed the task of measuring large fuselage components for most production processes. This requires the manual fixing of several reflector targets in specifically designed pre-fitting stations, which not only is time consuming but also prone to human error. Additionally, the measurement times tend to be very high due to the sequential measurement of targets. To address this challenge, this paper proposes a multi-camera system for marker-less measurement of the shape and position of large fuselage panels through the detection of its own features. This approach aims to improve the current fuselage section assembly by reducing its runtime, hardware costs and improving flexibility. Moreover, the integration of such a system is analyzed by utilizing it in shape adjustment experiments.

Keywords: Large-scale metrology · Aircraft manufacturing · Digital image processing · Camera calibration

1 Introduction and State of the Art

In the particular case of large fuselage panels, providing a sufficiently large measurement volume with a vision-based system that also meets the required process accuracy is the main challenge to address. Furthermore, to exploit the advantages of camera-based metrology systems in comparison to a laser tracker, it is important, instead of using markers, to rely on the detection of the components' own features.

Incorporating several stereo-camera systems and combining their respective measurement volumes through a common coordinate system makes it possible to expand the overall measurement volume and therefore, meet the required size. Linking the stereo-camera systems through overlapping parts in their individual measurement volumes results in unfavorable error propagation, which increases with each new stereo-camera introduced in the system. Furthermore, the individual measurement volumes would need to be relatively large to ensure a meaningful overlap. The solution proposed in this work

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 194–204, 2023.
https://doi.org/10.1007/978-3-031-16281-7_19

addresses the aforementioned problem by utilizing a laser tracker as a calibration reference tool while aiming the individual stereo-camera systems at local structural features of the outer fuselage without any overlap. For this purpose, the first step is to design a calibration body that can be detected by both laser tracker and camera systems. In addition, the concept and setup require the calibration method to differ from more common ones, as e.g. [1–3], that predominantly need estimating the pose of the calibration body to the cameras for each acquired image pair.

1.1 Aircraft Section Assembly

A typical long-range commercial aircraft is built from several semi-cylindrical sections. For some aircraft models, these sections are built according to a four-panel concept: After being separately manufactured, upper-, lower-, left- and right-panels are connected together with a floor grid. For positioning, forming and joining them along their longitudinal edges, a system of heavy-duty actuators with three degrees of freedom with grippers is used. One of such assembly systems currently in operation in the program of the Airbus A350 aircraft is shown in the left image in Fig. 1 while the right image roughly summarizes the steps for section assembly.

1.2 Large Fuselage Panel Metrology

Due to the size of the aircraft sections and the relatively tight assembly tolerances of around 0.4 mm, optical non-contact measurement systems such as laser trackers are widespread in this type of assembly process [4]. Before the assembly process starts, several spherically mounted retroreflectors (SMRs) must be installed at a pre-fitting station on each panel. Afterwards, during the assembly process, the targets are measured in sequence automatically. After the measurement routine, the assembly program compares the measured and the nominal coordinates and calculates the deformation movements for the grippers. This procedure is then repeated several times for one part until the deviation between the measured and nominal data is below a certain threshold while the introduced forces are monitored. The schematics in the middle image of Fig. 1 show a representation of a major component assembly system with a laser tracker.

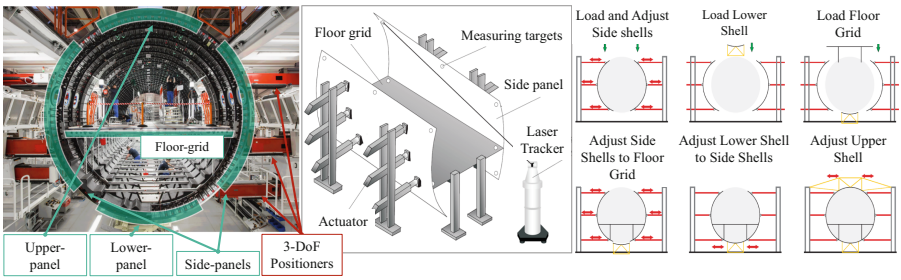


Fig. 1. Left: major-component-assembly station for section assembly of the A350 XWB fuselage ©Premium AEROTEC. Middle: positioning system and structure assembly of aircraft sections with laser tracker [5]. Right: general steps of the major-component assembly process [6]

2 Multi-camera System Calibration

For a process-conform three-dimensional measurement of the structural features in the panels, the parameters of each stereo-camera system must be determined with sufficient accuracy. This is done by comparing the real, experimentally measured system behavior with that of a model and parametrizing it in such a way that the deviation between both is minimal. Since the multi-camera system consists of multiple stereo-camera systems, its parametric model is defined so that it is composed of the intrinsic and extrinsic stereo-camera parameters of each stereo-camera system as well as their geometrical relationship to the reference coordinate system. A particular advantage of using a laser tracker as a calibration reference over common stereo-camera calibration strategies is that it is not necessary to estimate the pose of the calibration body in respect to the stereo-camera system for each individual calibration pose. Furthermore, a numerical rank analysis of the linear approximation to the camera model gives a new perspective on the sensitivity and interactions of the parameters, which improves calibration results.

2.1 System Setup

The calibration setup includes a multi-camera system itself, consisting of four camera pairs (IDS: GV-5200SE) with their corresponding lenses, lighting systems and filters, an industrial KUKA KR240 serial kinematic robot on a linear axis, which is used to position the calibration body and together, an AT401 laser tracker and a custom calibration body serve as the calibration reference. The camera systems are fixed each at the base of one panel-handling hexapod robot as seen in Fig. 2. Their orientation and focus is set up in accordance with the panel features.

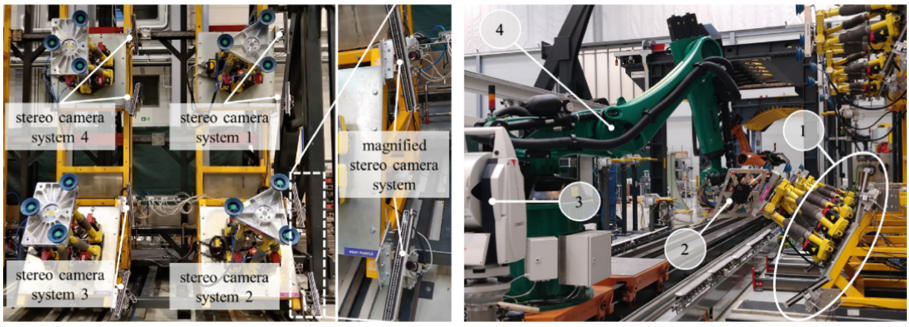


Fig. 2. Left: Multi-camera metrology system, integrated in the prototype facility for shape & position adjustment and magnified stereo-camera system. Right: Stereo-camera system calibration set-up consisting of: 1) Stereo-camera system, 2) Calibration body with calibration targets, 3) Laser tracker, 4) Actuator

At that, SMRs spread across on different parts on the floor of the facility define the global reference frame for the entire installation. The calibration body consists of eight precision bores with corresponding sphere mounts. The arrangement of the measuring

marks among the mounts was such that the retroreflective targets are mounted in the four middle locations and in three of the four mounts on the sides, three spherical mirror reflectors were placed. In Fig. 3, the calibration body and a diagram of the setup of the measuring marks can be seen.

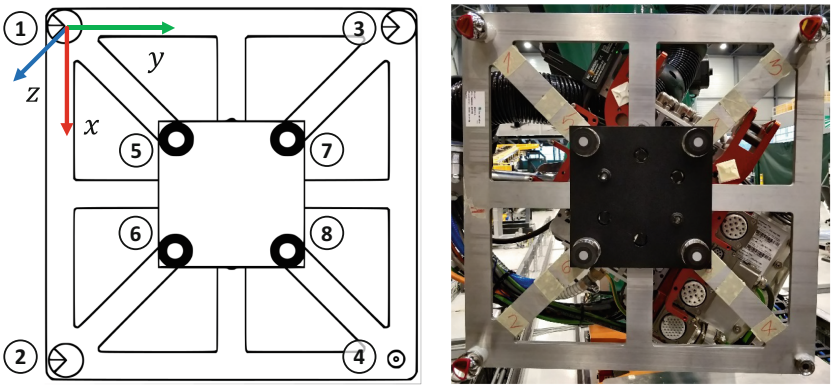


Fig. 3. Left: Diagram of SMRs (1–4) and retroreflective targets (5–8) placement on the calibration body plus coordinate system defined by SMRs. Right: Real robot-held calibration body with targets

Since the retroreflective targets and SMRs are interchangeable, the coordinates of the center points of the target mounts on the calibration body are measured beforehand with the laser tracker. Therefore, their relationship is known before calibration starts.

2.2 Calibration Method

The multi-camera model used in this approach is based on a standard stereo-camera model extended by the poses of the primary cameras in respect to the global reference. By using interchangeable mirror and retroreflective targets, the co-optimization of the parameters that describe the reference between SMRs and retroreflective targets is also avoided. Thus, the parameters to be determined for each stereo pair in the setup are shown in Eq. (1).

$$\mathbf{p}_{\mathrm{C0}} = \left(\mathbf{p}_{\mathrm{R}}^{\mathrm{C0}^{\mathrm{t}}} \; \mathbf{p}_{\mathrm{CI0}}{}^{\mathrm{t}} \right)^{\mathrm{t}}$$

$$\mathbf{p}_{\mathrm{C1}} = \left(\mathbf{p}_{\mathrm{C0}}^{\mathrm{C1}^{\mathrm{t}}} \; \mathbf{p}_{\mathrm{CI1}}{}^{\mathrm{t}} \right)^{\mathrm{t}}$$

$$\mathbf{p}_{\mathrm{sys}} = \left(\mathbf{p}_{\mathrm{C0}}{}^{\mathrm{t}} \; \mathbf{p}_{\mathrm{C1}}{}^{\mathrm{t}} \right)^{\mathrm{t}}$$

where:

$$\mathbf{p}_{\mathrm{R}}^{\mathrm{C0}} = \left(t_x^{\mathrm{C0}} \; t_y^{\mathrm{C0}} \; t_z^{\mathrm{C0}} \; \theta_x^{\mathrm{C0}} \; \theta_y^{\mathrm{C0}} \; \theta_z^{\mathrm{C0}} \right)^{\mathrm{t}}$$

$$\mathbf{p}_{\mathrm{C0}}^{\mathrm{C1}} = \left(t_x^{\mathrm{C1}} \; t_y^{\mathrm{C1}} \; t_z^{\mathrm{C1}} \; \theta_x^{\mathrm{C1}} \; \theta_y^{\mathrm{C1}} \; \theta_z^{\mathrm{C1}} \right)^{\mathrm{t}}$$

$$\mathbf{p}_{\mathrm{CI}_k} = \begin{pmatrix} b & \chi & \alpha_x & \alpha_y & S_x & S_y & H_x & H_y \end{pmatrix}_k^{\mathrm{t}} \tag{1}$$

The vectors $\mathbf{p}_{\mathrm{C0}}$ and $\mathbf{p}_{\mathrm{C1}}$ are the parameter vectors of the primary and secondary camera respectively, where $\mathbf{p}_{\mathrm{sys}}$ represents the sensor model for the complete stereo-camera system. These vectors consist of the extrinsic and intrinsic camera parameters that are the pose from secondary to primary camera $\mathbf{p}_{\mathrm{C0}}^{\mathrm{C1}}$, the pose from the primary camera to the global reference frame $\mathbf{p}_{\mathrm{R}}^{\mathrm{C0}}$ and their respective intrinsic parameters $\mathbf{p}_{\mathrm{CI0}}$ and $\mathbf{p}_{\mathrm{CI1}}$.

The spatial coordinates of the calibration points are calculated by extracting and undistorting their associated subpixel 2D-coordinates from a corresponding image pair. From this, rays of sight are calculated whose approximated intersection-point is calibration point in the stereo-camera coordinate system. These coordinates are then transformed to the global coordinate system as depicted in Eq. (2).

$$\begin{pmatrix} {}^{\mathrm{R}}\mathbf{r}_{ip} \\ 1 \end{pmatrix} = \mathbf{T}\left(\mathbf{p}_{\mathrm{R}}^{\mathrm{C0}}\right) \begin{pmatrix} {}^{\mathrm{C0}}\mathbf{r}_{ip}\left(\mathbf{p}_{\mathrm{sys}}, {}^{\mathrm{S}}\mathbf{r}_{ip}\right) \\ 1 \end{pmatrix} \tag{2}$$

where ${}^{\mathrm{C0}}\mathbf{r}_{ip}\left(\mathbf{p}_{\mathrm{sys}}, {}^{\mathrm{S}}\mathbf{r}_{ip}\right)$ are the 3D-coordinates of the calibration target in the stereo-camera coordinate system represented as a function of the sensor model $\mathbf{p}_{\mathrm{sys}}$ and the extracted, undistorted 2D-coordinates ${}^{\mathrm{S}}\mathbf{r}_{ip} = \left(\begin{pmatrix} x_i & y_i \end{pmatrix}_0 \begin{pmatrix} x_i & y_i \end{pmatrix}_1 \right)_p^{\mathrm{t}}$. The indices i and p describe the i-th calibration target in the calibration body and the p-th calibration pose, respectively. The matrix $\mathbf{T}\left(\mathbf{p}_{\mathrm{R}}^{\mathrm{C0}}\right)$ represents the homogeneous transformation correspondent to the pose $\mathbf{p}_{\mathrm{R}}^{\mathrm{C0}}$. The vector ${}^{\mathrm{R}}\mathbf{r}_{ip}$ represents the predicted coordinates and is depicted in the global reference system. The estimation of $\mathbf{p}_{\mathrm{sys}}$ for a single stereo-camera system is done by using the objective function $Q\left(\mathbf{p}_{\mathrm{sys}}\right)$, which depicts the root mean square deviation between the predicted coordinates ${}^{\mathrm{R}}\mathbf{r}_{ip}$ and the coordinates obtained through measurement with the laser tracker ${}_{meas}{}^{\mathrm{R}}\mathbf{r}_{ip}$ in dependence of $\mathbf{p}_{\mathrm{sys}}$ as shown in Eq. (3).

$$Q\left(\mathbf{p}_{\mathrm{sys}}\right) = \sqrt{\frac{1}{PI}\sum_{p=0}^{P}\sum_{i=0}^{I}\left\| {}_{meas}{}^{\mathrm{R}}\mathbf{r}_{ip} - {}^{\mathrm{R}}\mathbf{r}_{ip}\left(\mathbf{p}_{\mathrm{sys}}, {}^{\mathrm{S}}\mathbf{r}_{ip}\right)\right\|^2} \tag{3}$$

Through the numerical minimization of $Q^2\left(\mathbf{p}_{\mathrm{sys}}\right)$ for each camera pair, the optimal parameters for the entire multi-camera system are obtained. Since the calibration of each stereo system can be performed independently, the multi-camera system can be expanded with relative simplicity and flexibility. In addition, the accuracy of the laser tracker is still the upper limit for the accuracy of the whole measurement system within the entire measurement volume.

2.3 Data Acquisition

For the data acquisition, an equidistant grid with 32 measurement locations in front of each stereo-camera pair is defined. Each location-grid covers a measuring volume of about the size of an aircraft window (400 x 200 x 100 mm^3). The serial kinematic robot moves the calibration body to each of these locations, after which the stereo-camera system acquires an image pair. The subpixel coordinates ${}^{S}\mathbf{r}_{ip}$ of each retroreflective target in the image pair is then extracted while the laser tracker measures the SMRs and extrapolates the coordinates of the reflector targets ${}_{meas}{}^{R}\mathbf{r}_{ip}$.

2.4 Numerical Rank Analysis

To address the interdependencies between the parameters in the system, a numerical rank analysis was performed by simplifying Eq. (2) as a linear approximation at the initial estimate $\mathbf{p}_0$ for the model parameters. Thus, in the sense of linearization, the rank of the resulting matrix $\mathbf{J_p}(\mathbf{p}_0)$ corresponds to the number of linearly independent parameters of the model. Similarly, the eigenvalues of the matrix provide information regarding the order in which parameters should be identified based on their influence and interdependency [7]. The results of the numerical rank analysis in Table 1 are listed in the decreasing order of singular values.

Table 1. Resulting eigenvalues and corresponding camera parameters for a representative stereo-camera system

Parameter	Unit	Eigenvalue	Parameter	Unit	Eigenvalue
S_x^{C1}	m	7,61E+05	t_y^{C1}	m	2,52E−02
S_x^{C0}	m	1,80E+05	t_y^{C0}	m	1,31E−02
S_y^{C1}	m	3,97E+04	α_y^{C0}	rad	1,15E−02
S_y^{C0}	m	2,89E +04	θ_x^{C0}	rad	7,77E−03
b^{C1}	m	2,86E+01	θ_y^{C0}	rad	6,59E−03
b^{C0}	m	2,38E+01	α_x^{C1}	rad	2,11E−03
θ_y^{C1}	rad	1,84E+01	θ_z^{C0}	rad	1,62E−03
θ_z^{C1}	rad	1,27E+00	$\theta_x^{C_1}$	rad	1,11E−03
t_x^{C1}	m	4,85E−01	$\boldsymbol{H_x^{C_0}}$	**pixel**	**8,79E−06**
α_x^{C0}	rad	3,70E−01	$\boldsymbol{H_y^{C_0}}$	**pixel**	**2,44E−06**
α_y^{C1}	rad	2,66E−01	$\boldsymbol{\chi^{C_0}}$	-	**1,33E−07**
t_x^{C0}	m	1,95E−01	$\boldsymbol{H_x^{C_1}}$	**pixel**	**9,69E−08**
t_z^{C1}	m	1,11E−01	$\boldsymbol{H_y^{C_1}}$	**pixel**	**4,48E−08**
			$\boldsymbol{\chi^{C_1}}$	-	**8,73E−09**

The radial distortion coefficients χ^{C_0}, χ^{C_1} and the principal points $H_x^{C_0}$, $H_y^{C_0}$, $H_x^{C_1}$, $H_y^{C_1}$ parameters are not identified as their singular values are too small to make a meaningful change in the resulting coordinates. Subsequently, a parameter-optimized calibration sequence was created based on the results of the numerical rank analysis and personal expertise. After the definition of the sequence, the minimization of the objective function is performed using conjugate directions as in [8].

2.5 Calibration Result

The results of three different calibration routines for one representative stereo-camera systems are shown in Table 2.

Table 2. Residuum of the calibration process for a stereo system with three different datasets

Residuum	Data sets				
	unit	1	2	3	σ
Q	m	7,97E−05	8,05E−05	7,95E−05	4,25E−07
r	pixel	0,173	0,172	0,171	0,001

Since the residuum from the calibration process can only be used as an indicator of the overall measurement accuracy, a validation test using the same setup as that of the calibration process was performed. The validation is based on the guidelines for verification of optical measurement systems found in the VDI 2634 Part 1 [9]. While the data is acquired in the same manner as in calibration, the poses the calibration body takes are different so that the validation and calibration data sets are different. Additionally, the validation was performed two weeks after the calibration taking into account environmental influences as vibration and temperature shifts in a realistic industrial environment. The validation results for all the four stereo-camera systems are listed in Table 3.

Table 3. Mean residual after the validation process for all camera systems

Mean Residual	Camera system				
	unit	#1	#2	#3	#4
Q	m	1,36E−04	9,70E−05	9,20E−05	1,83E−04
r	pixel	0,289	0,309	0,171	0,283

Thus, the overall accuracy of the system can be roughly summarized as 0.127 mm $\pm$ 0.037 mm, which is within the tolerance for the assembly of large aircraft structures.

3 Shell Deformation and Shape Adjustment

To analyze the metrology system in an operational environment, shape adjustment experiments were performed with a ca. 5 x 5 m CFRP upper-panel segment. Four parallel kinematics Fanuc F-200ib robots with end-effectors equipped with vacuum-grippers and force & torque cells act as the panel-handling system as seen in Fig. 2 and Fig. 4.

3.1 Measurement of Rivet Groups as Features and Setup

The rivet heads on the coated panel surface serve as reference marks by means of which the geometry of the part is defined, measured and the correction movements of the manipulators is calculated. The riveting of the panels is done with tight tolerances during panel manufacturing and the rivets themselves are available in the CAD model of the aircraft. Thus, they are well suited to act as features in the process. To obtain the rivet coordinates, an algorithm was developed that sorts the rivets in the images into regions of interest (ROI), calculates their subpixel coordinates by extracting the rivet contours and fits them to ellipses. Subsequently, the ellipse-center pairs are projected onto the 3D-space as in [7]. The setup of the stereo-camera corresponding to the second manipulator and a sample of the acquired images can be seen in Fig. 4.

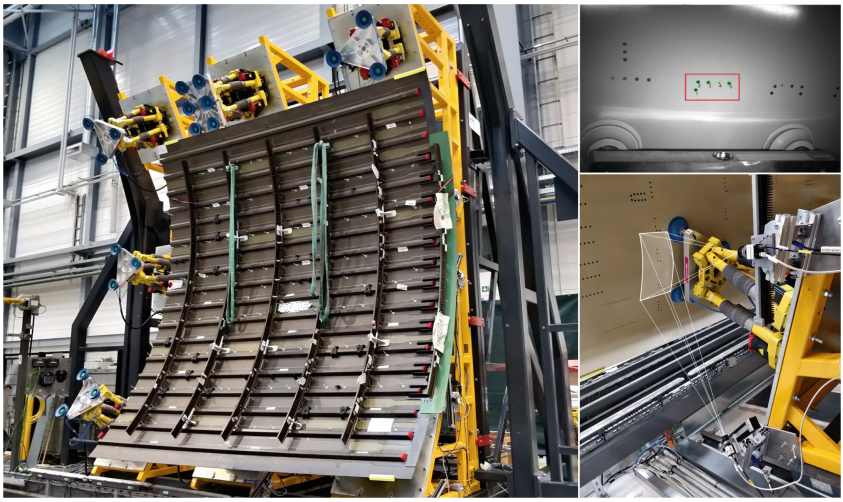


Fig. 4. Left: CFRP upper-panel on manipulator system, Right: stereo-camera setup with field of view and resulting stereo images with rivet group and corresponding ROIs

The measurement process eliminates the need for mounting markers and therefore reduces the production run time by several hours. Likewise, the measurement time for an entire panel is reduced from several minutes to a few seconds.

3.2 Deformation and Iterative Shape Adjustment

To avoid potentially damaging the assembly system or a panel during the adjustment processes, measured applied forces and torques must be factored in and predicted as part of the iterative process. For this, the assumption was made that for small deformations, a simplified linear model of the interactions between actuator and rivet position as well as measured forces as in [6] would suffice.

After manually mounting the panel, the shape adjustment process starts by measuring the rivet coordinates and comparing them to the corresponding target coordinates in the CAD shell model. A best-fit procedure is used to calculate optimal intermediate target rivet coordinates, which define the correct shape of the shell while minimizing the path of movement of the robot. This process is iteratively performed until the deviation for the rivets is below a threshold or the estimated forces for the next step surpass a given limit. The results of one of the iterative shape adjustment tests for the upper-panel for all manipulators can be seen in Fig. 5.

The graphs in Fig. 5 show that the deviation decreases with each adjustment step while the applied forces and torques generally increase. In this case, the procedure was stopped before the rivets reached their target to avoid potentially damage to the manipulators by the introduced forces.

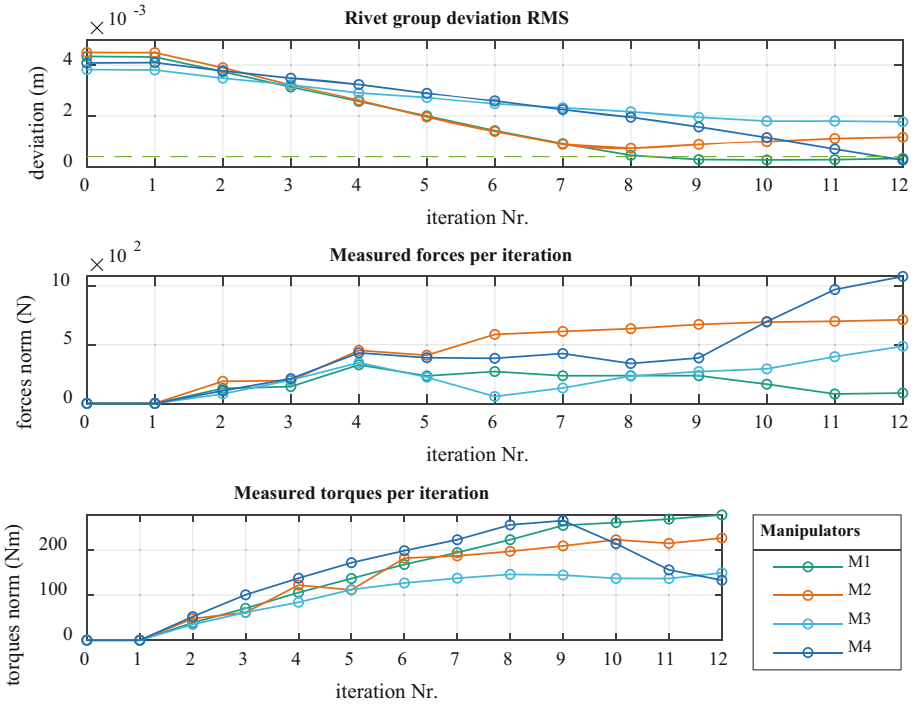


Fig. 5. Rivet position deviation, measured forces and torques per correction step

4 Conclusion and Outlook

The result of this work presents a successful implementation of an alternative, vision-based measurement strategy for the assembly of large aircraft panels. Through the application of a specific calibration method and reference, it was possible to exceed the required accuracy of the standard aircraft assembly process as shown in Table 3. The multi-camera system can be simply expanded in order to measure larger components without much loss of accuracy, since e.g. the AT401 loses only about 6 µm per meter [10]. Likewise, due to its flexibility, the system can be customized to address particular challenges of different assembly systems as varying production tolerances and measurement volumes. Moreover, the proposed system measures an entire panel in the time a typical laser tracker needs to measure a single SMR in production today. The test results shown in Fig. 5 demonstrate that while the algorithm reduces the deviation from the target geometry, the number of manipulators was insufficient to deal with the required forces. Likewise, the stagnation of forces may indicate minimal slippage of the vacuum gripper after certain shear forces may have been exceeded. This can be addressed by increasing the number of manipulators and using ones with higher load capacity. Further, immediate next steps that would improve the assembly process are:

- Utilization of sophisticated panel structure and force prediction models, for example the FEM method for large components presented in [11].
- Inclusion of deep learning and neural networks to shape adjustment as in [12].
- Use of structured lighting to measure panel interfaces.

Considering the next steps, this work offers great potential to significantly reduce costs by reducing production time, eliminating pre-fitting stations, replacing costly hardware and increasing flexibility in large structure assembly.

Acknowledgement. The presented investigations were conducted within the project BiSconA (LuFoV3-FKZ: 20W1724B), funded by the Federal Ministry for Economic Affairs and Climate Action (Project partners: Airbus Operations GmbH, BA Assembly & Turnkey Systems GmbH, CENIT AG, Fraunhofer IFAM, 3D.aero GmbH). The authors thank Dr. Dirk Niermann (head of the department) and the project partners for their support and advice.

References

1. Zhang, Z.: A flexible new technique for camera calibration. IEEE Trans Pattern Anal Machine Intell **22**, 1330–1334 (2000). https://doi.org/10.1109/34.888718
2. Steger, C., Ulrich, M., Wiedemann, C.: Machine vision algorithms and applications. In: 4th repr. Textbook. Wiley-VCH, Weinheim (2015)
3. Luhmann, T.: Nahbereichsphotogrammetrie: Grundlagen, Methoden und Anwendungen, 3., völlig neu bearb. und erw. Aufl. Wichmann, Berlin (2010)
4. Cuypers, W., van Gestel, N., Voet, A., et al.: Optical measurement techniques for mobile and large-scale dimensional metrology. Opt. Lasers Eng. **47**, 292–300 (2009). https://doi.org/10.1016/j.optlaseng.2008.03.013

5. Stepanek, P.: Flexibel automatisierte Montage von leicht verformbaren großvolumigen Bauteilen. Schriftenreihe des Arbeitsbereichs Werkzeugmaschinen und Automatisierungstechnik der Technischen Universität Hamburg-Harburg, vol 16. Shaker, Aachen (2007)
6. Ramirez Martinez, J.A.: Flexible automated assembly systems for large CFRP structures using geometrical and force information. Dissertation, Technische Universität Hamburg-Harburg (2015)
7. Wollnack, J.: Videometrische Verfahren zur Genauigkeitssteigerung von Industrierobotern. Zugl.: Hamburg-Harburg, Techn. Univ., Forschungsschwerpunkt 5, Werkstoffe - Konstruktion - Fertigung, Habil.-Schr., 2000. Berichte aus der Robotik. Shaker, Aachen (2001)
8. Powell, M.J.D.: An efficient method for finding the minimum of a function of several variables without calculating derivatives. Comput. J. **7**, 155–162 (1964). https://doi.org/10.1093/comjnl/7.2.155
9. VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik Optical 3D measuring systems - Imaging systems with point-by-point probing: VDI/VDE 2634 Part 1(2634:2022–05) (2022)
10. Leica Geosystems AG: Leica AT401 User Manual, Switzerland (2010)
11. Bertelsmeier, F.: Einsatz eines Cyber-Physischen Produktionssystems zur flexiblen Positionierung und Ausrichtung von dünnwandigen Großbauteilen, 1. Auflage. Fertigungsmesstechnik et Qualitätsmanagement, Band 14/2019. Apprimus Verlag, Aachen (2019)
12. Neuhaus, F., Schildt, R., Juedes, W., et al.: Production system for the automated assembly of vehicle components and method for controlling a production system. EP20160171696 B64F5/00;G05B19/402;G05B19/418 (EP3249482 (A1)) (2016)

Management of Research Field Data Within the Concept of Digital Twin

Hauke Dierend(✉), Osman Altun, Iryna Mozgova, and Roland Lachmayer

Institute of Product Development, Leibniz University Hannover,
30823 Garbsen, Germany
dierend@ipeg.uni-hannover.de
https://www.ipeg.uni-hannover.de/

Abstract. In engineering research, various sensors and data sources are utilized for data acquisition. Together with subsequent processing and analysis, a large, heterogeneous amount of data and information is generated. To structure data and information within the research project, intelligent tools for research data management (RDM) are required. However, existing tools for RDM focus on data management during a research project and lack of capabilities to describe a technical system over this life cycle in multiple projects. Thus this paper addresses the potential of the Digital Twin (DT) concept for RDM. Based on requirements from RDM we derive specifications for a DT concept in RDM and introduce three criteria to choose suitable technical systems for DT. We present a DT for a research vehicle, implemented within the open-source Industry 4.0 DT tool "AASX Package Explorer", to show the benefits of using a DT for RDM.

Keywords: Digital Twin · Research data management · Field data

1 Introduction

Nowadays, various measurement systems and data sources, e.g. public datasets or web services, are used for data collection in engineering research. The heterogeneity of research data ranges from measurement and simulation data to literature sources and images for documentation purposes. The efficient management of data and information generated in research processes and the extraction of knowledge are critical factors for the success of research projects and publications. Intelligent tools and methods for research DM are required to enhance data findability, accessibility, interoperability, and reusability according to the FAIR Data Principles [25]. This affects data management within a research project, publication of data, or their subsequent use for follow-up projects. In particular, the structured and machine-readable description of technical systems used and the linking of data and models are hardly addressed by previous tools in research data management [2]. The concept of the Digital Twin (DT), a digital replica of a physical asset with data interfaces between the two, promises to

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 205–214, 2023.
https://doi.org/10.1007/978-3-031-16281-7_20

solve this challenge. The DT has been state of the art for several years and offers researchers an integrated environment for data management along the life cycle of a technical system. Although a DT is mainly introduced to enable simulating the physical twin, this paper focuses on structuring and managing data with the DT as it provides the foundation for simulations at a later stage of the life cycle. Therefore, this paper first introduces the concept of the DT. Chapter 3 presents a concept for DTs in research data management and defines the criteria for choosing suitable technical systems for implementation. Based on the concept and criteria, Chap. 4 presents an exemplary implementation of a DT for a test vehicle using the tool "AASX Package Explorer".

2 Data Management

Data Management (DM) includes a range of aspects and processes related to the collection, modification, storage, and retrieval of data. It should be noted that the literature, depending on context and purpose, considers approaches to DM in research and DM in development and production. These two contexts cannot be separated from each other, but they do lead to different methodological and software approaches and solutions used in educational, research, and commercial organizations.

2.1 Research Data Management

Research data management (RDM) is the process of transforming, selecting, and storing data generated during scientific activity to keep it accessible, reusable and verifiable in the long term and independently of the data creator. RDM intervenes at every stage of the research data life cycle to preserve the significance and accessibility for other researchers and to secure the chain of evidence. The quality assessment of one's own RDM can be carried out using the FAIR Data Principles, according to whose guidelines data has to be Findable, Accessible, Interoperable, and Reusable [25]. Table 1 provides an overview of the criteria. As researchers increasingly rely on computer tools to cope with the enormous amount of heterogeneous data, the principles emphasize the machine-readability of data. When discussing data, however, the context in which it was generated, as well as its meaning and purpose have to be provided [14].

Table 1. Requirements to fulfill FAIR data principles according to [25]

	Description
Findable	Data have rich metadata and a unique identifier
Accessible	(Meta)data can be retrieved by identifier using standardized protocols
Interoperable	(Meta)data use an accessible and standardized language
Reusable	(Meta)data are well described and contain license information

Research founders in the national and international space have introduced guidelines for the RDM to make research data and results a public good [5,7]. The proposed German National Research Data Infrastructure (NFDI) aims to bring together existing infrastructure components and services for a comprehensive, interdisciplinary and sustainable RDM. The NFDI4Ing consortium represents engineering science and community requirements [19].

Amorim et al. give an overview of RDM systems to support the RDM workflow [2]. In the scientific community, the use of open networked DM systems is preferred when selecting data management concepts and tools. Systems based on the Comprehensive Knowledge Archive Network (CKAN) are widely used for archiving research data. Open-configurable APIs, the creation of domain-specific vocabularies, and the expansion of services through plugins are all supported by CKAN. As a result, the service may be extended with tools for executing software scripts or displaying research data [1,21]. Various electronic lab book solutions or wiki-based systems are also popular for documenting research results [14]. However, at the moment, these systems do not provide a direct opportunity to link data sources and models with associated research data.

2.2 Product Data Management and Product Life Cycle Management

Product Data Management (PDM) and Product life cycle Management (PLM) systems are used in the industrial world to arrange data storage and management. Within the development process, PDM systems are used to manage computer-generated models and other documentation. PDM is typically regarded as an important aspect of PLM because it extends capabilities to a product's life cycle perspective [23]. A PDM system aids in the development of technical systems by ensuring that all parties involved have access to the same information. In the development phase, modern PDM systems assist the processing, analysis, archiving, and later reuse of system-relevant data. PLM systems generally document the industrial processes of product and service engineering. However, uniform interfaces and protocols for accessing and using data are still needed [18].

Modern PLM systems, as well as FDM systems, support the processing, analysis, archiving and subsequent reuse of product-relevant data in the development phase. However, for the planning and simulation of components or processes, there is only limited consistency in the product life cycle. Monitoring during application is also not fully supported by an available PLM approach. An important aspect of DM is the completeness and correctness of the description of domain-specific data, which cannot be achieved without a culture of using thesauri, vocabularies and classifiers of data. Vocabularies and classifiers are used to describe the content of objects and add significant value to metadata, as does the use of a standardized schema. The FAIR Data Principles provide clear requirements for metadata and their properties for DM. In addition, there are typical requirements regarding good scientific practice, e.g. reference to the work of other researchers. Further analysis shows that only very few of DMSs

are actually specialized in engineering data, in terms of metadata schemes or supported data types. The same holds for the Research Data Alliance Metadata Standards Directory, listing not a single standard or tool for the engineering sciences, except for crystallography [1]. Both mentioned aspects - the RDM for technical products or systems over their life cycle explicitly and the completeness of data description - lead us to the need to further search for forms and systems of DM. As such a form we propose to consider the concept of a Digital Twin. Therefore, it is essential to identify requirements for applying a DT in RDM as well as providing criteria for researchers to decide, if a DT fits their needs.

2.3 Digital Twin

The first definition of a DT goes back to a presentation on product life cycle management by Grieves at the University of Michigan in 2003, according to which a DT consists of a physical product, a digital representation of this product, and interfaces for the exchange of data between the two [9]. Later Grieves et al. further linked the DT to the product life cycle by introducing digital representations for the prototypical physical product (DT Prototype), a specific physical product through the entire product life cycle (DT Instance), and the environmental operating conditions impacting the product life (DT Environment) [10]. In addition to Grieves definitions, there are others: According to Stark et al., the DT can be understood both as a real product or a product-service system that contains "selected characteristics, properties, states and behaviors by means of models, information and data within a single or also over several life cycle phases" [22]. Glaessgen and Stargel define the DT as "an integrated multi-physics, multi-scale, probabilistic simulation of a vehicle or system in its as-built state that uses the best available physical models, sensor updates, fleet history, etc. to map the life of the corresponding flying twin" [8]. Inherent in all definitions is that a DT represents a physical entity, adapts its models using real data from the field, and thus realizes an accurate description of the physical entity along with its life.

The concept of the DT is being researched and applied in multiple areas. It is most widely utilized in manufacturing, where field data is utilized for process optimisation and simulation of factory plants [13]. For predictive maintenance, real-world data is used to simulate future entity behavior and improve service intervals or detect anomalies early [4]. In civil engineering, DTs enhance the architectural modeling of single buildings up to whole cities [16]. In the automotive domain, Wang et al. connect DTs of vehicle, driver, and traffic in the cloud to enable personalized advanced driver-assistance systems [24].

The connection of models and data, and thus the traceability of the technical system, offer the greatest potential for applying the concept of DT. The uniform data structure in the DT and the direct reference to the technical system simplify for engineers to get started with DM. In addition to the bare data, a system description based on models is essential for traceability and subsequent use of data. Through domain-specific, standardized metadata and globally unique identifiers, contained in a structured and machine-readable format, system understanding can be facilitated for reuses.

3 Digital Twin in Research Data Management

3.1 Concept of a Digital Twin for RDM

The DT structures and stores data and information over the life cycle of a technical system, which is the basis and data source of research in engineering. Figure 1 shows the interaction between research projects, a physical twin, and its DT. RDM offers the greatest benefit when it is applied from begin of the development of a technical system to its end. Therefore, a DT must be created in the design phase, when a physical twin does not yet exist. The DT receives data and information on research results in the literature, on data sets, and models and algorithms from other researchers or its projects. In line with good scientific practice to distinguish achievements of others from one's own and at the same time to implicitly enable documentation of the source, metadata describing the origin of the approaches and tools must be provided and included in the DT. Standardized metadata for citing research publications, e.g. the Data Cite MetadataSchema, have to be applied [3]. Static data, e.g. publications, must be declared with global identifiers like the Digital Object Identifier (DOI). Dynamic data, e.g. public datasets or websites, must be tagged with version numbers and access dates.

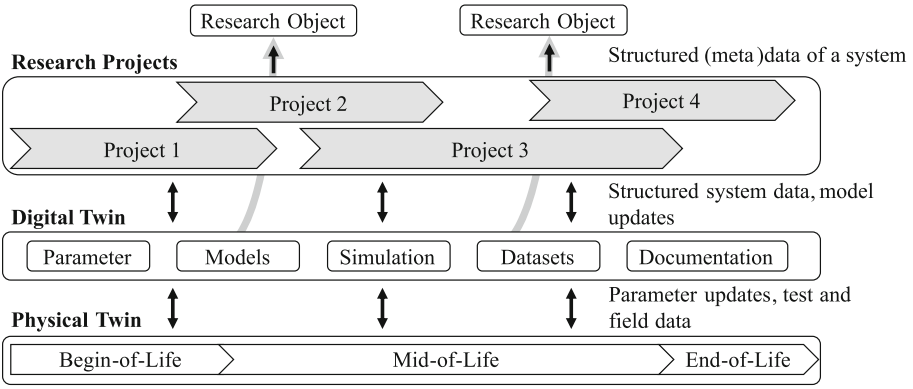


Fig. 1. Interaction between research projects, DT, and physical twin during the life cycle of the latter

During the construction of the technical system, information about its elements is integrated into the DT or further DTs of subsystems are created, which in turn are linked to the DT of the overall system. Translation errors or the use of terms that are not clearly defined can limit the interpretability of metadata for humans and machines. Common Data Dictionaries such as IEC 61630 offer standardized and machine-readable terminology for technical systems and services. In addition to parameters for system description, e.g. the diameter of a

component, links to documentation pages such as a wiki or electronic lab books are integrated.

The technical system typically undergoes many tests and calibrations during its construction and in preparation for experimental operation. Key quality parameters, e.g. the signal-to-noise ratio, which can quickly give researchers and subsequent users an understanding of the system quality and thus the significance of the research results, should be explicitly listed in the DT. Further data and results of these tests are provided with a digital local identifier and their files are linked via a URL, for example. The technical system is put into an operation concerning the research question and the DT receives data from sensors and algorithms of the physical twin. The data can be received in the DT via an application programming interface (API) such as HTTP, OPC UA or MQTT or with a time delay, e.g. via a USB stick [11]. Due to the utilization of identifiers in data exchange, the former can directly assign the data to a model or an element of the DT. For the latter, this is done manually. The central requirement for the storage location is that the researcher can access data from the DT at any time. The data can be stored in the technical system, on the researcher's private computer, and/or on internal and external databases like MongoDB [17].

Analysis programs and algorithms can query (meta)data via the APIs of the DT. These include field data sets of sensors, operating parameters, and calibration data. The APIs also enable cooperative access for other researchers. Due to the structured storage of data and its description via metadata, the transfer into non-proprietary file formats such as XML, AutomationML, JSON, or CSV is possible [20]. These generic data formats allow downstream users to use the data without implementing a DT themselves. While the life of the physical twin ends at the end of the project or a few years after, the research data must be available for multiple years. E.g. the German Research Foundation requires primary data of a project to be archived for ten years [6]. Non-proprietary file-formats facilitate the transfer to other tools for long-term storage.

Researchers may only consider integrating metadata at the end of their research activity. Providing a DT for a technical system used in many research projects decouples metadata integration from single research activity. Research institutions will enforce researchers developing the technical system to insert metadata plus parameters and implement APIs early to provide seamless documentation and reduce the effort in following research activities. Furthermore, the DT, working as an interface between researcher and technical system, enables parameter updates that directly affect the system. This will encourage researchers to use the DT as part of their everyday research activities.

3.2 Derivation of Suitable Applications for Digital Twin in RDM

The DT concept can enhance DM for technical systems. However, the effort of creating a DT is not always worthwhile. Before implementation, the following three central criteria should therefore be analyzed by the researcher to assess the suitability of one or more DTs for the technical system. The criteria are derived

from the system perspective to enable engineers, as they are the developers and maintainers of the DT, the decision making:

- *Level of detail of the system decomposition*: A DT can represent a single system component of an overall system, e.g. a sensor, up to the overall system itself, e.g. a robot. The former enables a flexible exchange of system elements with little change over the life cycle and a high level of detail for system elements. However, implementation of each system and subsystem requires effort and time, defining interfaces, integrating standardized metadata descriptions, and linking models and data.
- *Scope of system information*: The researcher has to incorporate the requirements of potential downstream users as well as the own. For example, in a company, detailed information about manufacturers and suppliers of individual system elements may be required, whereas in academic research this information does not provide any added value for downstream users.
- *Diversity of system elements and interfaces*: The DT should include standardized metadata and global identifiers to comply with FAIR Data principles. Searching for appropriate metadata and identifiers takes a large part of implementation time. If many identical or similar elements are used, the effort is significantly reduced.

In addition, there are other aspects such as the availability of resources (time, money, knowledge) for integration and maintenance of the DT or the type of data exchange within the research project.

Based on the criteria applications can be identified for which the use of a DT is suitable. These include technical systems that are operated over several years and used for various research projects and publications. Here, the DT can show its strengths by providing complete documentation and data history about the system and also enabling the integration of models for predictive maintenance of the system. Another suitable application of the DT is in the combination of field data and synthetic data, e.g. in the development of sensor models for simulation. Short-lived setups of technical systems with many different components are an unsuitable application for the DT, because both descriptions of all components and the development of valid digital models are too time-consuming.

4 Exemplary Implementation of a Digital Twin for RDM

A large number of tools exist that focus mainly on industrial applications [15]. These include commercial tools such as Microsoft's Azure DTs or the open-source solutions AASX Explorer and BaSyx. The latter implement the Industry 4.0 standard. Here, a technical system is referred to as an asset and is digitally described by means of an Asset Administration Shell (AAS), a standardized DT. The syntax of the AAS is defined by a protocol-independent type system and meta-model. The AAS refers to sub models (SM) that provide services to implement operations and exploit the asset's data [12]. BaSyx contains Java, C#, and C++ SDKs to develop AASs, while the AASX Explorer provides a

Graphical User Interface (GUI) to create and edit AASs in XML and JSON. Thus, the AASX Explorer allows an easier start to get familiar with the AAS and, therefore, is used for the implementation of our use-case.

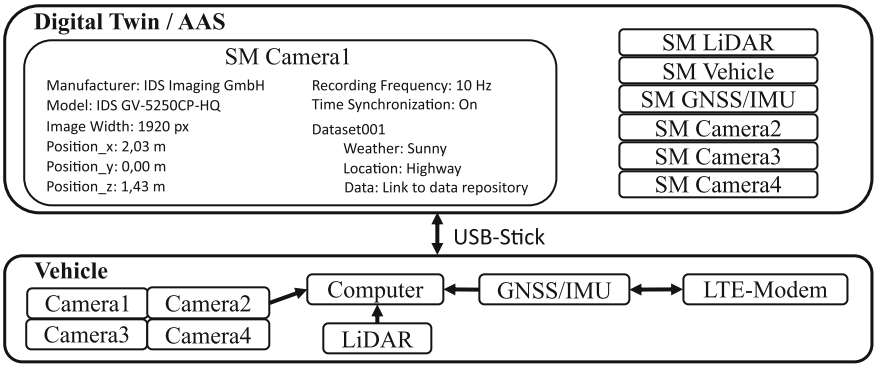


Fig. 2. Data and Information about a research vehicle managed in the AAS. Exemplary data and information of a camera sub module are shown.

Our technical system is a vehicle equipped with a computer and different sensors, such as LiDAR and cameras (see Fig. 2). The vehicle is represented by an AAS. Each component, e.g. a camera, is described by means of an SM. The SMs contain static and dynamic information. The former is information about the manufacturer and technical constants, e.g. the position of a component, as well as links to internal documentation on the institute's server or publications by other researchers that describe, for example, the methodology used for camera calibration. The publications are cited using the DataCite Metadata Schema. Data recordings from the vehicle and project-specific, changeable parameters, e.g. the recording frequency, represent dynamic information.

In terms of findability, the single data parts have a *semanticId* each, which enables unique, local, or global referencing, e.g. to IEC 61630. The *idShort* is used for naming according to metadata standards, e.g. the DataCite Metadata Schema. Through the two IDs and the description of the identifiers and standards used, the data can be found in the AASX Explorer and simultaneously realize interoperability and re-use between researchers and different research projects.

The AAS and all related information and files are stored in the XML-based .aasx file format. Since the AASX Explorer is an open-source tool and also provides conversion of SMs to AutomationML or JSON, the data is accessible to other researchers. However, before publishing, the researcher must ensure that links, e.g. to internal documents, are accessible to other researchers. Alternatively, they must explicitly provide the data and modify the AAS for publication.

The AASX Explorer does not directly address the re-use of data by providing licensing information. Although the AASX Explorer enables authentication and thus implicit licensing of data access, this is only possible for a"living" AAS. In

the case of publications, however, the researcher must explicitly document why some data are missing and which licenses belong to the data.

5 Conclusion

In this paper, we presented the DT as a concept for the DM of research field data. We showed the potential of DTs for representing a technical system during its life cycle, thus being used in different research projects to structure and manage data and information about the system. We implemented our use-case of a vehicle to demonstrate employing the AASX Explorer as a tool for a quick setup of a DT by integrating homogeneous field data with standardized metadata descriptions and identifiers and addressing the FAIR Data Principles.

However, an important characteristic of the DT, the simulation of the physical twin based on field data and models, has not been evaluated yet. Especially for larger systems, the modeling and simulation can increase the benefits to apply the DT concept using tools like AASX Explorer. Furthermore, ready-to-use templates can reduce implementation time for engineers, enhance the application of the tools and provide a standardized description within research communities.

Acknowledgement. The authors would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713.

References

1. Altun, O., et al.: Integration eines digitalen maschinenparks in ein forschungsdatenmanagementsystem. In: Proceedings of the 32nd Symposium Design for X (2021)
2. Amorim, R.C., Castro, J.A., Rocha da Silva, J., Ribeiro, C.: A comparison of research data management platforms: architecture, flexible metadata and interoperability. Univ. Access Inf. Soc. **16**(4), 851–862 (2016). https://doi.org/10.1007/s10209-016-0475-y
3. DataCite Metadata Working Group: Datacite metadata schema documentation for the publication and citation of research data and other research outputs v4.4 (2021)
4. Errandonea, I., Beltrán, S., Arrizabalaga, S.: Digital twin for maintenance: a literature review. Comput. Ind. **123**, 103316 (2020)
5. European Research Council: Guidelines on implementation of open access to scientific publications and research data (2019)
6. German Research Foundation: Allgemeine Bedingungen für Förderverträge mit der Deutschen Forschungsgemeinschaft e.V. (DFG) (2015)
7. German Research Foundation: Leitfaden für die Antragstellung **54**, 01 (2022)
8. Glaessgen, E., Stargel, D.: The digital twin paradigm for future NASA and u.s. air force vehicles. In: 53rd Structures, Structural Dynamics and Materials Conference. American Institute of Aeronautics and Astronautics (2012)

9. Grieves, M.: Digital twin: Manufacturing excellence through virtual factory replication (2014)
10. Grieves, M., Vickers, J.: Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In: Transdisciplinary Perspectives on Complex Systems, pp. 85–113 (2016)
11. Jacoby, M., Usländer, T.: Digital twin and internet of things—current standards landscape. Appl. Sci. **10**(18), 6519 (2020)
12. Kannoth, S., Hermann, J., Damm, M., Rübel, P., Rusin, D., Jacobi, M., Mittelsdorf, B., Kuhn, T., Antonino, P.O.: Enabling SMEs to industry 4.0 using the BaSyx middleware: A case study. In: Software Architecture, pp. 277–294 (2021)
13. Kritzinger, W., Karner, M., Traar, G., Henjes, J., Sihn, W.: Digital twin in manufacturing: A categorical literature review and classification. In: 16th IFAC Symposium on Information Control Problems in Manufacturing, vol. 51(11), pp. 1016–1022 (2018)
14. Mozgova, I., Koepler, O., Kraft, A., Lachmayer, R., Auer, S.: Research data management system for a large collaborative project. In: Proceedings of NordDesign (2020)
15. Qi, Q., et al.: Enabling technologies and tools for digital twin. J. Manuf. Syst. **58**, 3–21 (2021)
16. Ruohomaki, T., Airaksinen, E., Huuska, P., Kesaniemi, O., Martikka, M., Suomisto, J.: Smart city platform enabling digital twin. In: 2018 International Conference on Intelligent Systems (IS) (2018)
17. Sangat, P., Indrawan-Santiago, M., Taniar, D.: Sensor data management in the cloud: data storage, data ingestion, and data retrieval. Concurrency Comput. Pract. Exp. **30**(1), 1–10 (2017)
18. Scheidel, W., Mozgova, I., Lachmayer, R.: Structuring information in technical inheritance with pdm systems. In: Proceedings of the 21st International Conference on Engineering Design (ICED 2017), vol. 6, pp. 217–226 (2017)
19. Schmitt, R.H., Anthofer, V., Auer, S., Başkaya, S., Bischof, C., et al.: Nfdi4ing - the national research data infrastructure for engineering sciences
20. Schroeder, G.N., Steinmetz, C., Pereira, C.E., Espindola, D.B.: Digital twin data modeling with AutomationML and a communication methodology for data exchange. In: 4th IFAC Symposium on Telematics Applications, vol. 49, pp. 12–17 (2016)
21. Sheveleva, T., Koepler, O., Mozgova, I., Lachmayer, R., Auer, S.: Development of a domain-specific ontology to support research data management for the tailored forming technology. In: Proceedings of the 5th International Conference on System-Integrated Intelligence, vol. 52, pp. 107–112
22. Stark, R., Damerau, T.: Digital twin. In: CIRP Encyclopedia of Production Engineering, pp. 1–8 (2019)
23. Ulrich, T.: Datamanagement for production companies. In: PLM-Jahrbuch 2016 - Der Leitfaden für den PLM Markt, pp. 60–73
24. Wang, Z., Gupta, R., Han, K., Wang, H., Ganlath, A., Ammar, N., Tiwari, P.: Mobility digital twin: Concept, architecture, case study, and future challenges. IEEE Internet Things J. 1–1 (2022)
25. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3**(1), 160018 (2016)

Feed-Forward SNN for Touch Modality Prediction

Ali Dabbous[1], Ali Ibrahim[1,2(✉)], and Maurizio Valle[1(✉)]

[1] Department of Electrical, Electronic, Telecommunication Engineering and Naval Architecture DITEN, University of Genoa, Genova, Italy
AliDabbous@edu.unige.it, maurizio.valle@unige.it

[2] Department of Electrical and Electronics Engineering, Lebanese International University (LIU), Beirut 1105, Lebanon
ali.ibrahim@liu.edu.lb

Abstract. Recently, Spiking Neural Networks (SNNs) have been considered as alternatives to the common deep neural networks (DNNs) when the energy efficiency has been targeted. The SNNs adopt an event-driven information processing approach in which the computational expenses are reduced considerably compared to DNNs without affecting the system performance. This paper presents an efficient framework based on SNNs for touch modality classification. The proposed work outperforms similar state of the art solutions by achieving a classification accuracy of 99.97% with decreased complexity and increased number of classes.

Keywords: Tactile sensing system · Spiking neural network · Neuromorphic hardware · Brian2 simulator · Touch modality

1 Introduction

Recent technological advances have shifted the role of robots from simple industrial tools in controlled environments to intelligent devices that can assist humans and interact with the environment [1]. Several applications have emerged in the robotic domain attempting to mimic the human behavior such as humanoids and social robots, or to enable and restore difficult tasks in the case of medical robots and artificial limbs. This shift creates new challenges that must be addressed to ensure safe and efficient interaction between robots and humans while providing manufacturing, entertainment, education, healthcare, and healthcare services. A particular study in the field of neuromorphic engineering is dedicated to the emulation of the neuronal function and organization of the human nervous systems in electronic devices with the aim of advancing the efficiency and robustness of computational interactions with environments. Mimicking the communicative function of neurons action potentials through spikes that encode and propagate information, represents an example of such emulations [2]. In addition to neuromorphic hardware, event-driven sensors are modeled based on biological models to perform the function of dealing with stimuli from the environment as spikes.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 215–222, 2023.
https://doi.org/10.1007/978-3-031-16281-7_21

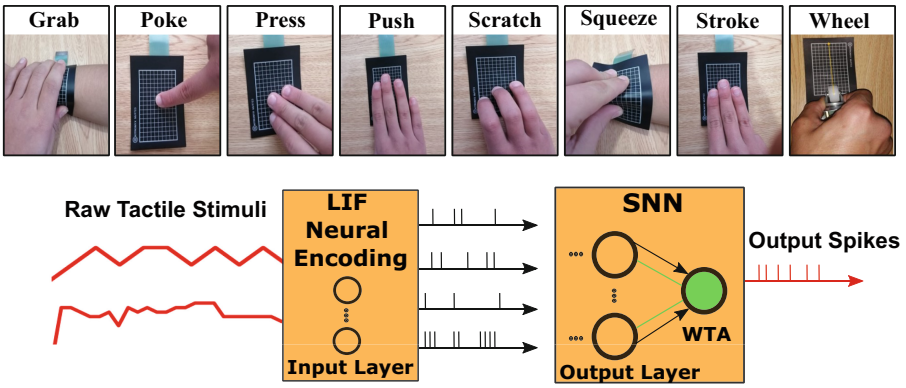


Fig. 1. Touch modality architecture; (Top) eight touch modalities were exercised on the sensory array from 3 participants. (Bottom) 2 layers SNN for touch modality classification; input layer consists of 160 neurons eight output neurons (single neuron for each modality), in which the connection between the two layer were all-to-all connection through weighted synapses. The network include also Winner-take-All competition represented by global inhibitory neuron for better selectivity.

Beside sensors, bio-inspired computational structures including spiking neural networks have been created [3]. Modeling the intricate dynamics of spiking neurons allows the implementation of biological learning and plasticity mechanisms. For improved realism, Spiking Neural Network (SNN), which are synthetic neural networks, mimic the rich time-varying dynamics observed in biology [3,4], and they outperform rate-based neural network models for complicated tasks like recognizing visual patterns, smells, sound qualities, and tactile stimuli [5,6]. SNNs are excellent for simulating information processing in the brain and researching diverse learning mechanisms because of their biological credibility [7,8]. Pattern recognition and real-time computation are two other engineering issues that benefit from the utilization of SNNs [9,10]. Synaptic connections connect the neurons in a SNN network. Through synapses with variable weights or efficacies, every higher-order neuron can take information from several lower-order neurons.

Computational neuron models in SNNs, like their biological counterparts, interact by spikes, that represent discrete time events. The capacity of neurons to transmit particular spatio-temporal input spike patterns to temporally accurate output spikes is very beneficial for SNNs with multiple layers, as the spiking outputs can subsequently be decoded by higher-order neurons which have sensitivity to spike timing [11]. The weights of the input synapses could be tuned to obtain temporally precise mapping. Moreover, a neuromorphic approach of integrating the hardware sensory system along with spiking neuron models can compute considerably more quickly than conventional computers, while using significantly lower power. Therefore, the use of neuromorphic concepts and hardware will revolutionize the world of tactile sensing. Several recent studies have concentrated on the classification/recognition of materials shapes,

textures and patterns [12,13]. In [14], authors used a SNN consist of three layers of spiking neurons in order to recognise between various bar orientations exercised on artificial skin. In [15], authors proposed a two layers SNN for object shape recognition. Nevertheless, the identification of gestures and touch modalities is addressed in fewer studies [16–19]. In [20], authors used KNN to classify 6 touch modalities with a total accuracy equals to 77. In [21], authors used 3 SNN layers to classify 6 touch modalities by adjusting the synaptic weights between the hidden and output layers using Spike Timing Dependent Plasticity (STDP). The results show that the proposed system gives good performance in classifying different touch modalities with total accuracy of 88.3%. However, because of the complexity of spike trains, developing effective procedures for the computation weights remains a difficulty. In this paper, the implementation of a neuromorphic SNN model for the classification of 8 different touch modalities is presented. The network model consists of 2 layers. A piezoresisitive sensory array have been used where the resulted pressure values have been encoded into spike trains using the (LIF) neuron model. The contributions of the paper may be summarized as follows:

- An overall computationally efficient touch modality classification framework able to make fast inference and represents a possible candidate to be run on neuromorphic chips such as the Spinnaker;
- An efficient and simple architecture composed of only two layers neuromorphic spiking neural network tested on tactile dataset to classify different touch modalities;
- A supervised synaptic learning rule STDP is implemented enabling to learn the network differentiating the different modalities;
- The proposed system outperforms similar state of the art solution in [21] by achieving a 99.97% accuracy while having a decreased number of layers (2 layers compared to 3 layers) and eight touch modalities to be classified instead of six in [20,21];

2 Methods

2.1 Dataset

In this work, we used the piezo-resistive sensory array in order to obtain a dataset consisting of 8 different touch modalities as shown in the Fig. 1. The tactile sensory array has 160 force sensing nodes. Three participants (two male and one female) with an average age of 20 were cooperated to collect the touch modalities dataset. When one participant applied one of the touch modalities on the sensory array, the electronic interface (snowboard system) detects the generated forces signal and sends them to the computer by the use of a USB interface. Snowforce 3 software has been used in this work in order to visualize and test the system, as well as to collect data for the touch modality dataset. Each participant was instructed to place their finger on the sensor array constituting the modality for 1 to 3 s at a time. The touch modality dataset is divided into eight classes, each

with 90 samples. The samples were divided in a random way into 63 samples representing the training set (70 %) and 27 samples for the testing set for each class (30 %).

2.2 Two Layers Spiking Neural Network

For tactile signal classification, we use a two-layer feedforward SNN. The network is composed of only one input layer and one output layer. The input layer composed of 160 neurons that encodes the input tactile signal to trains of spike, and 8 neurons (single neuron per touch modality) which decodes the input touch modality, composes the output layer (Fig. 1-bottom). Each node in the sensory array is connected to only a single neuron from input layer as one to one connection, whereas neurons in input and output layers were connected as all to all connection. In this paper, the LIF neural model has been used for the input and output layer neurons, simulated with Brian2 simulator [22]. the dynamic membrane potential u in this model is described in Eq. 1

$$\frac{du}{dt} = \frac{u_{rest} - u + R_m * I(t)}{\tau_m} \tag{1}$$

$$\text{if} \quad u_{mem}(t) > u_{th}. \quad \text{then} \quad S(t) \leftarrow 1; \quad u_{mem}(t) \leftarrow u_{reset}$$

In this equation, $I(t)$ represents the injected current to the neuron, R_m is the membrane Resistance and τ_m is the membrane time constant. When applying an input current, the membrane potential u_{mem} of the neuron starts to increase until reaching a fixed threshold u_{th}. When the neuron reaches its threshold, if fires a spike and then goes back to its resting value u_{reset}. S represents the output spike train of the neuron. It is a function whose value is 1 when the neuron fires a spike at time t and 0 otherwise. For input layer neurons $I(t)$ represents the output tactile signals generated from the sensors when the participants applied each of the touch modalities.

2.3 Synapse Model

Synaptic weights connecting output and input layers neurons were changes according to supervised STDP. As an inspiration by the work of [21], the changing of synaptic weights connecting input and output layer neurons depends on the post-synaptic spike traces, post-synaptic target traces and input layer neurons spikes as shown in Eq. 2.

$$w_{ij} = \alpha.(tgt_i^{post} - tr_i^{post})\delta(t - t_j^k) \tag{2}$$

in which w_{ij} are the dynamic synaptic weights linking the i_{th} neuron of the output layer to the j_{th} neuron in the input layer, $alpha$ is the learning rate, tgt_i represents the post synaptic target trace constant, $delta(t_{jk})$ is the time needed for the jth neuron representing the input layer to fire a spike. In addition, the

post-synaptic spike traces is modified every time an output layer neuron fires a spike ($\delta(t - t_i^k)$) and decreases exponentially over time (τ_{tr}), Eq. 3

$$\tau_{tr} \frac{dtr_i}{dt} = -tr_i + \sum_k \delta(t - t_i^k) \tag{3}$$

The weights changed at every input layer neuron fires a spike, in which following Eq. 2 they will be potentiated or depressed.

3 Simulation Results

3.1 Synaptic Weights Learning

The aim of this work was to learn the various touch modalities applied by the participants on the sensory array by using only two layers (input and output layers). We used supervised STDP learning between the input and output layers using all to all connections to achieve this goal. At the start, all of the weights were equal to zero. The output layer is fed by several stimulus presentations consecutively throughout the learning phase, in which the eight different touch modalities were repeated randomly. The activity of spiking for a number of neurons in the input layer encodes each stimulus presentation. Moreover, the learning dataset used was consist of 504 samples (8 modalities per 63 each). In addition, at each stimulus presentation, we fed one neuron of the output layer with external current, in which $I(t)$ in Eq. 1 was equal to an array of current in order for the neuron to fire a spike. For instance, when the grab modality was presented to the network we fed only neuron number zero in the output layer with an external current that acts as a teacher signal. Therefore the change of the weights (depression or potentiation) based on combining the input layer of the neurons activated and the teaching signal.

Each curve in Fig. 2 represents the summation of potentiated weights connecting neurons in the input layer to single neuron in the output layer. Depending on the activated neurons of the input layer for each modality, a specific connectivity pattern is guaranteed for each neuron in the output layer. After learning, synaptic weights between input and output layer neurons alter gradually, resulting in persistent connection patterns. Therefore, thanks to STDP learning and the teaching signals the network has the can to learn the eight various touch modalities.

3.2 Touch Modality Classification

The weights extracted from the learning phase are used to test the ability of the network in predicting the correct input touch modality. In addition, we switched off the learning contributions. During testing, the current I(t) in Eq. 1 equals to

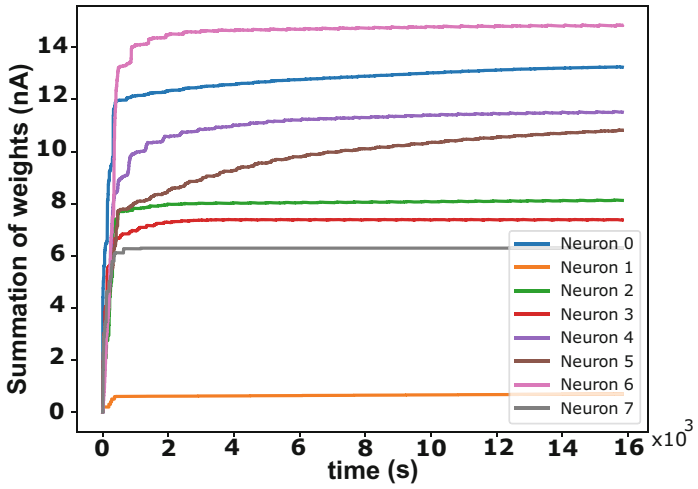


Fig. 2. Summation of weights showing the connection of the neurons in input layer to the neurons in the output layer during learning.

the summation of weights at every input neurons fires a spike, and are written in the Eq. 4 below:

$$I_i(t) = \sum_{j=1}^{n} \left(w_{ji} \sum_{k} Sj\,(t - t_k) \right) + I_{inh} \tag{4}$$

S_j are the spatio-temporal output spikes of the j_{th} neuron in input layer. t_k represents the time needed for the neurons in input layer to fire a spike. To achieve better accuracy during testing, WTA has been implemented represented by a global inhibitory neurons, which receive excitations from neurons in the output layer and in then it send an inhibitory current for all of them (I_{inh}). In the testing phase, each of the touch modality presented to the network for 27 different samples, which is equal to 30% of the total dataset for each modality.

Figure 3 shows the confusion matrix after testing. It is clearly seen that the six modalities: "Poke", "Press", "Push", "Scratch", "Squeeze", and "Wheel" can be identified accurately by our network with 100 %, whereas the "Grab" and "Stroke" modalities achieved 99.92% and 99.88% accuracy, respectively. These findings show that the proposed network is efficient, with an accuracy of 99.97 %, outperforming state-of-the-art work [20,21].

	Grab	Poke	Press	Push	Scratch	Squeeze	Stroke	Wheel	Accuracy
Grab	25	0	0	0	0	0	2	0	99.92 %
Poke	0	27	0	0	0	0	0	0	100 %
press	0	0	27	0	0	0	0	0	100 %
Push	0	0	0	27	0	0	0	0	100 %
Scratch	0	0	0	0	27	0	0	0	100 %
Squeeze	0	0	0	0	0	27	0	0	100 %
Stroke	3	0	0	0	0	0	24	0	99.88 %
Wheel	0	0	0	0	0	0	0	27	100 %

Fig. 3. Touch modality classification Confusion matrix .

4 Conclusion

This work presented an SNN architecture for classification of various touch modalities. The architecture is based on two layers of LIF neurons decreasing the computational complexity of the network. In order to provide the network to learn and differentiate among various touch modalities, the work adopted a supervised synaptic learning rule STDP between the input and the output layers. The results show that the proposed network is efficient, with a classification accuracy of 99.97%, outperforming equivalent state-of-the-art methods. Future work will consist of the implementation of the proposed architecture on neuromorphic chips and/or the deployment of this methods on energy efficient hardware devices to be integrated near to the sensors location enabling different application domains such as prosthetics and humanoid robotics.

References

1. Dahiya, R.S., Metta, G., Valle, M., Sandini, G.: Tactile sensing-from humans to humanoids. IEEE Trans. Rob. **26**(1), 1–20 (2009)
2. VanRullen, R., Guyonneau, R., Thorpe, S.J.: Spike times make sense. Trends Neurosci. **28**(1), 1–4 (2005)
3. Maass, W.: Networks of spiking neurons: the third generation of neural network models. Neural Netw. **10**(9), 1659–1671 (1997)
4. Gütig, R., Sompolinsky, H.: The tempotron: a neuron that learns spike timing-based decisions. Nat. Neurosci. **9**(3), 420–428 (2006)
5. Bologna, L.L., Pinoteau, J., Brasselet, R., Maggiali, M., Arleo, A.: Encoding/decoding of first and second order tactile afferents in a neurorobotic application. J. Physiol. Paris **105**(1–3), 25–35 (2011)
6. Gütig, R., Gollisch, T., Sompolinsky, H., Meister, M.: Computing complex visual features with retinal spike times. PLoS ONE **8**(1), e53063 (2013)
7. Nessler, B., Pfeiffer, M., Maass, W.: Stdp enables spiking neurons to detect hidden causes of their inputs. Adv. Neural. Inf. Process. Syst. **22**, 1357–1365 (2009)

8. Legenstein, R., Pecevski, D., Maass, W.: A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. PLoS Comput. Biol. **4**(10), e1000180 (2008)
9. Schrauwen, B., D'Haene, M., Verstraeten, D., Van Campenhout, J.: Compact hardware liquid state machines on fpga for real-time speech recognition. Neural Netw. **21**(2–3), 511–523 (2008)
10. Rice, K.L., Bhuiyan, M.A., Taha, T.M., Vutsinas, C.N., Smith, M.C.: Fpga implementation of izhikevich spiking neural networks for character recognition. In: 2009 International Conference on Reconfigurable Computing and FPGAs, pp. 451–456. IEEE (2009)
11. Gütig, R.: To spike, or when to spike? Curr. Opin. Neurobiol. **25**, 134–139 (2014)
12. Talbot, M., Arvandi, M., Sadeghian, A.: A neural network based surface roughness discrimination algorithm. In: 2008 World Automation Congress, pp. 1–8. IEEE (2008)
13. Decherchi, S., Gastaldo, P., Dahiya, R.S., Valle, M., Zunino, R.: Tactile-data classification of contact materials using computational intelligence. IEEE Trans. Rob. **27**(3), 635–639 (2011)
14. Dabbous, A., Mastella, M., Natarajan, A., Chicca, E., Valle, M., Bartolozzi, C.: Artificial bio-inspired tactile receptive fields for edge orientation classification. In: 2021 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5. IEEE (2021)
15. Kim, J., et al.: Object shape recognition using tactile sensor arrays by a spiking neural network with unsupervised learning. In: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 178–183. IEEE (2020)
16. Iwata, H., Sugano, S.: Human-robot-contact-state identification based on tactile recognition. IEEE Trans. Industr. Electron. **52**(6), 1468–1477 (2005)
17. Stiehl, W.D., Breazeal, C.: Applying a "somatic alphabet" approach to inferring orientation, motion, and direction in clusters of force sensing resistors. In: 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No. 04CH37566), vol. 3, pp. 3015–3020. IEEE (2004)
18. Gianoglio, C., Ragusa, E., Zunino, R., Valle, M.: 1-d convolutional neural networks for touch modalities classification. In: 2021 28th IEEE International Conference on Electronics, Circuits, and Systems (ICECS), pp. 1–6. IEEE (2021)
19. Gianoglio, C., Ragusa, E., Gastaldo, P., Valle, M.: A novel learning strategy for the trade-off between accuracy and computational cost: a touch modalities classification case study. IEEE Sens. J. **22**(1), 659–670 (2021)
20. Sun, J., Billing, E., Seoane, F., Zhou, B., Högberg, D., Hemeren, P.: Categories of touch: classifying human touch using a soft tactile sensor. In: The robotic sense of touch: From sensing to understanding, workshop at the IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May 2017 (2017)
21. Dabbous, A., Ibrahim, A., Valle, M., Bartolozzi, C.: Touch modality classification using spiking neural networks and supervised-stdp learning. In: 2021 28th IEEE International Conference on Electronics, Circuits, and Systems (ICECS), pp. 1–4. IEEE
22. Stimberg, M., Brette, R., Goodman, D.F.: Brian 2, an intuitive and efficient neural simulator. eLife **8**, e47314 (2019)

Smart Factory and Logistic Systems

Ansaldo Energia Progetto LHP (OR6.3)

Proper Management of PPE (Personal Protective Equipment) Financed by the Italian Ministry of Economic Development

Roberto Mosca[1], Marco Mosca[1(✉)], Roberto Revetria[1], Saverio Pagano[2], and Federico Briatore[1]

[1] Mechanical, Industrial and Transport Engineer Department (D.I.M.E.), University of Genoa, 16126 Genoa, Ge, Italy
mosca@diptem.unige.it, marcotulliomosca@gmail.com, roberto.revetria@unige.it
[2] Ing. CEO Smart Track Srl., Lausanne, Switzerland
Saverio.Pagano@smarttrack.it

Abstract. Ansaldo Energia is a Major Player in Italy for metal mechanic production which decided to adopt Industry 4.0 standards. One of the main projects, at this purpose, was to develop a new vision concerning Safety. The goal was achieved by a Team built specifically by the Company. Particular importance was assigned by the Team to PPE (Personal Protective Equipment), devices designed to improve safety of the Operators in carrying out their duties. The problem that the Team has clearly warned, given the frequency of occurrence, concerns the periodic maintenance of PPE, to be carried out by law for each device, according to precise rules in order to preserve their efficiency as well as safety certification. Therefore, the Team studied a new methodology, the subject of this paper, based on hardware and software tools designed to monitor the legal revisions and critical deadlines of each PPE in use.

Keywords: PPE maintenance · PPE IOT · PPE risk · PPE 4.0

1 Introduction

As part of the LHP (Light House Plant) project co-financed by MISE (Italian Ministry for Economic Development) for the transformation of Ansaldo Energia into an Industry 4.0 factory, great importance was given to Sect. 6, regarding the safety of operators. The DIME (University of Genoa Department, literally "Dipartimento di Ingegneria Meccanica, Energetica, Gestionale e dei Trasporti"), was in charged as technical and scientific advisor of the project. From a first screening of the literature of the sector, the authors realized the gravity of the safety problem at a global level. As evidence of this consideration, the data provided by the ILO-International Labor Organization, report that worldwide every 15 s an operator dies and 153 are injured, with an annual projection of 1.2 million deaths and 320 million accidents recorded with a consequent overall economic loss of the order of 225 billion USD. The enormity of these data from the economic side, but

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 225–234, 2023.
https://doi.org/10.1007/978-3-031-16281-7_22

even more so from the human side, pushed the team to make efforts to make a positive contribution. The LHP project thus became the starting point for conceptualizing and implementing effective solutions. The subject of this paper is the correct maintenance of PPEs (Personal Protective Equipment), which are vital if properly revised and maintained at the deadlines (indicated by the manufacturer and made mandatory by law). Ansaldo and UNIGE (DIME), in this regard, decided to select a technological partner, with specific skills and to build up a multidisciplinary project team to tackle the problem and to conceptualize a new 4.0 system.

2 Literature Review (State of the Art)

A careful literature review was conducted by researching predominantly on Scopus, WOS and Google Scholar. Over 32.800 papers were found by researching about PPE, filtering initially by "PPE maintenance" and "smart PPE". The majority of the papers came from Scholar, nevertheless, the most were out of scope. In fact, soon, the authors realized that "smart PPE" was a misleading keyword, as referring to special applications, being not relevant to the current study. Therefore, a first selection was made by restricting to "PPE maintenance", reducing the paper considered to 17.391. It was then realized that the keyword "PPE maintenance" was used for a wide range of applications, for this reason the topic was focused on Industry, obtaining a further restriction to 299 papers only. The relevance of the articles obtained was higher, but still too much generic. Another filter was then applied, to concentrate the research on the goals of the study, by introducing new keywords like "risk assessment, PPE 4.0, health monitoring, IOT, wearable devices, electronic devices", shortening the list to 86 paper, which were deeply analyzed, further decreasing the selection to 23 papers, considered totally in line with the goals of this work. This result made the authors aware that literature does not currently treat much about the industrial automatic monitoring of PPE, to improve their maintenance and management. The literature relating to PPE frames the devices under different applications (concerning, for example, the use, the sector, the materials, the way to wear them [1, 4, 5, 7–13, 15, 17, 19, 23],..), nevertheless only a part of the same, including that part of literature published by the bodies responsible for the safety of the operators, underlines the great importance of a management of PPE that guarantees: a) efficiency over time, through compliance with the necessary maintenance; b) decommissioning on the expiry dates indicated by the manufacturers.

Inefficient PPEs become, in fact, useless and therefore dangerous for those who feel protected, with consequent risks for the safety of the operators. In this regard, in Italy, INAIL (National Institute for Accident Insurance) in 2019 published the [17], in which it is reiterated that for any device (PPEs included) it is through maintenance that the preservation of performance and safety features is guaranteed over time. Italian legislative decree 81/2008 in article 78 obliges workers who use PPE to take care of their maintenance. The ISO 14119 standard provides the possible combination between PPE and RFID (Radio Frequency IDentification) and dictates some restrictions on maintenance [3, 22]. In this last article are also included the benefits of the combination of PPE and RFID, a topic which is covered again by the article [18]. The papers [15, 20] emphasize the importance of the maintenance of the PPEs that protect the respiratory

tract of the operators. This is because, in the event of a malfunction of the protective equipment, they are led to tamper with it, taking all the consequences for their safety. Commonly used technology for such purposes is the IOT (Internet of Things). The selection reported in the bibliography, mentioned at the beginning of this chapter, cites the extensive literature consulted by the authors on the subject, which highlights what is of specific interest for this paper, in relation to the importance of a correct PPE supervision of dates established for maintenance and legal revisions, and the definitive expiry dates of the same, which concludes their life cycle [2, 3, 6, 13, 16, 19]. The literature review allowed the authors, already convinced by the investigation conducted in the company of the need to create a methodology to monitor the scheduled interventions to keep PPEs in efficiency, to concentrate: a) on devising an appropriate management model of such equipment; b) on conducting an accurate benchmark among the available technologies, to identify the most appropriate one.

3 Case Study

3.1 Team Building

After careful examination, Smart Track S.r.l. was selected as most suitable technological partner because of the great competences in 4.0 tracking technologies. Then the extended team was defined:

- Ansaldo Energia S.p.a. (EHS: Eng. Andrea Magro, Coordinator - Eng. Massimiliano Ziveri, Project Leader – ICT: Eng. Stefano Santucci, Head; Eng. Sergio Botti, System Engineer – DATT: Mr. D. Furfaro, PPE Responsible);
- Genoa University - DIME (Prof. Eng. Roberto Revetria, PhD: Scientific Coordinator, Prof. Eng. Marco Mosca, PhD: Technological e Scientific Advisor);
- Smart Track S.r.l. (Technological Partner: Eng. Saverio Pagano, AD, Eng. Simone Peirani, Coordinator, Eng. Alessandro Cortese, PM);
- Futuro S.r.l. (PPE Supplier, Ds. Margherita Pitto, Owner, Mr. Paolo Priano, Interfaccia A.E.).

3.2 Scenario AS-IS (Current Methodology)

To date, an internal function manages the PPEs manually, by means of registers, on which the manager records personal data, deliveries and withdrawals of new, stored or assigned devices. Despite his efficiency, it is easy to understand the objective difficulties inherent in this type of activity, based on “many to many” relationships (like manufacturers, PPE, operators, sites of use, …).

3.3 Scenario TO-BE (Innovative Methodology Proposed by the Team)

The activity involves the installation of appropriate sensors on a selection of PPEs associated to an acoustic and visual detection systems, to monitor the scheduled dates both for the periodic revisions and for the expiration dates. The installation of these

sensors will prevent any delay in control tasks, thus ensuring PPEs correct maintenance, revision or replacement within the established dates. Also, it will help keeping the operators aware of the importance for safety, of the correct management and maintenance of protection devices. At this point, the team proposed the developed methodology to the manager responsible for PPE. Once agreed on the objectives of this activity, the reference technology was selected, together with the technological partner (Smart Track). All the PPEs in use in the company were carefully assessed and the most suitable ones to start a pilot project were identified. The team therefore launched a specific research project; in the first phase of investigation, some elements of importance emerged, such as the fact that PPEs are subject to legal revisions and do not have an indefinite duration, but a precise life cycle. It also clearly emerged that monitoring and maintaining PPEs requires work, therefore, it is complex and time consuming. Another aspect to take into account is that the assignment of PPEs to operators traveling to construction sites, implies a correct estimate of the duration of the trip and the preventive calculation of revisions/expiration dates. Omission of this practice would expose the company and the operators to risk on safety.

3.4 Benchmarking of Technologies

Next step of the project was, for each technology identified, (like bar code, RFID and NFC-Near Field Communication) to highlight the strengths and weaknesses in relation to the specific use. Bar code was a consolidated and low-cost technology. On the other hand, they are not suitable for application on many types of surfaces such as lanyards or any PPE exposed to water; moreover, an uncomfortable reading of the same if placed internally and the ease of deterioration of coded labels. For this reason, RFID were considered, with numerous points in favor including the robustness of the system, the greater amount of information that can be stored on 96bit tags and the possibility of making multiple scanning contemporary. Also, the RFID tags are more compatible than bar codes with the geometry of the selected PPEs. Another point in favor is that these tags, sourced in volumes, are really cheap (in the order of hundredths of Euro). The weakness of this solution, for this specific application, is that the gates to be monitored are many, in fact, the operators can access the workplace from different directions (enter and exit from different doors of the warehouses) and can insist on different work islands both in the factory and at the construction sites. This layout would imply a redundant cost for antennas (the antennas required range from € 600 to € 1000 each). The last technology investigated was NFC, with some favorable characteristics like the practicality of application, since this type of tag can be integrated into the head of an electrician cable tie. Another strength consists in the ease of reading and the fact that NFC does not require expensive reading hardware, contrary to RFID antennas. In fact, an App for NFC smartphones already enables functions like scanning of the tags and access to the database. Weaknesses of this technology, compared to RFID, are the need to scan every single PPE and the cost of 1,00€ for each tag. At the conclusion of this benchmark analysis, the team selected, for this type of application, the technology offering the best ratio price vs. performance, still guaranteeing effective and efficient performance, represented clearly by NFC.

3.5 Feasibility and Sustainability Study

Regarding this phase of the project, the team needed to review what emerged from the previous together with ICT (Information & Communication Technology is the internal function managing both the contracts with technical suppliers and the systems installed in the factory). Two other review were held respectively with the DATT (internal function managing and assigning PPEs to operators) and with the external partner supplying PPEs (Futuro S.r.l.), with the scope to present the selection of PPEs identified for the pilot project. In this regard, the team proposed: a) harnesses; b) lanyards; c) electrically ventilated helmets; d) dielectric gloves; e) reels; f) belts. This selection was re-evaluated by the supplier to confirm the suitability. The supplier, as well, contacted each PPE manufacturer to get a written confirmation that the application of the tags cannot be considered as tampering. In this regard, the manufacturers contact, in turn, the relevant certification bodies, to ensure that the application does not cause the decay of the certifications. The outcome of the inquiry made it possible to ascertain that all the selection made was compatible with the integration of the NFC tags, exception made for the dielectric gloves which, if pierced to pass the cable tie housing the TAG, would lose the ability to isolate. The team, then, discharged such glows from the selection and proceeded with a re-evaluation of the remaining, in the field. This phase began with a meeting towards DATT, together with Futuro, to test, from an operational point of view, the application of the tags to each approved PPEs. The authors then proceeded to identify the best application for each PPE, meaning that it should not interfere with the work of the operators, nor generate annoyance, nor be exposed to shocks or sliding. The next steps consisted in: a) design of the operating process; b) design of the smartphone application.

3.6 Framework of the Project Steps

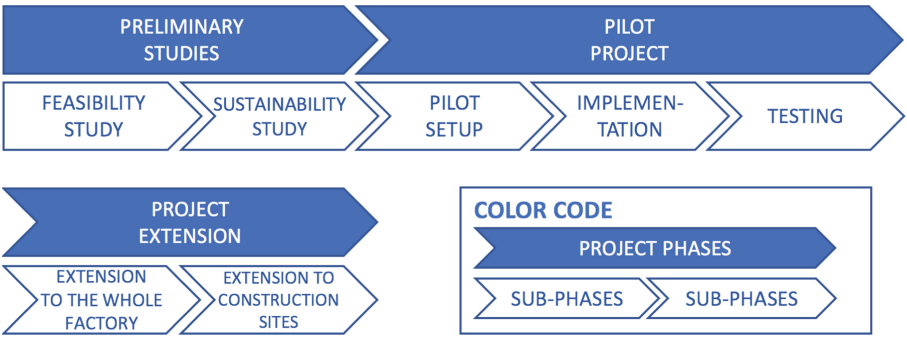


Fig. 1. Process of implementation

The framework proposed in Fig. 1 is applicable on the following conditions: a) full sponsorship of the management, providing a to the project the necessary resources (people and budget); b) creation of an internal function dedicated to the management of PPEs; c) census of PPEs available in the facility (type, supplier, brand, batch number,

serial number, expiry date, …); d) selection of the PPEs to be monitored; e) creation of a register for delivery and return of PPEs assigned to operators; f) adoption of the system (server, software, tags, NFC smartphone); g) integration of the tags on the PPEs; h) training to the involved personnel.

3.7 Design of the Operational Process

The authors then, produced clear photographs for each tag applied, showing the integration with PPEs, to get the approval by the Manufacturer. A new element, fundamental for the correct management of PPEs was to define a reporting procedure to track the equipment along the assignment. Such procedure is divided in 4 steps, appropriately noted in dedicated books: a) written request for PPE by the operator; b) approval by the competent office; c) delivery to the Operator; d) return of the PPE to the office. A similar procedure is used for the replacement of a PPE.

3.8 IIOT (Industrial Internet of Things)

The proposed system acts in compliance with Industry 4.0 standards, adopting the principles of IIOT, like the ability to generate a network of physical objects (not native for Internet connection, properly the selected PPEs), by embedding them with electronics and software, enabling Cloud connection and data exchange. The System, therefore, allows monitoring and controlling functions by a multitude of heterogeneous peripherals such as computers (fixed and mobile), tablets and smartphones. This comes through the designed application or through direct access from the web portal (only for authorized devices, in accordance with cyber security). The impact of the IIOT on industry generates multiple benefits. In this case, starting from safety, but with the possibility to develop a multitude of applications (like proactive maintenance, predictive maintenance, energy saving, monitoring-control-automation of the infrastructure in real time, integration of heterogeneous devices and of different manufacturers, environmental monitoring,…). The possibility of centralizing the data, collected at the peripheral level, also allows the development of a high analytical capacity, leading to significant improvement of the process knowledge. Considered the significant number of fatal accidents that occurs every day in industrial plants all over the world, this system represents an asset of a certain importance against both the threats on the safety of the Operators and the economic damage for companies.

3.9 Application for Smartphones

The diagram (Fig. 2) represents the architecture of the system, including the smartphone application, created with the aim of exploiting the most of the available technologies offered by a market, which is increasingly based on Industry 4.0.

In fact, many of the modern smartphones include NFC technology, which allows the acquisition of information by the appropriate tags, through the direct scanning procedure described above. The application communicates with a cloud hosted on a company server, which records the data on separate channels in a relational database. The analytics

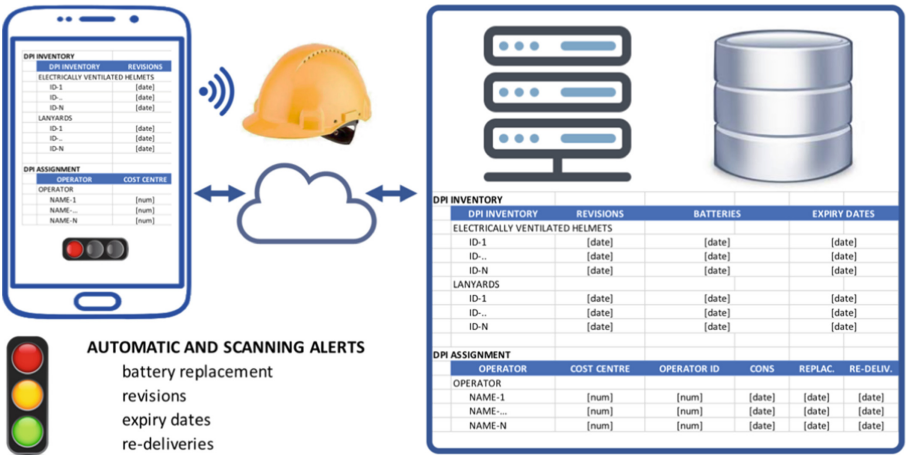


Fig. 2. System architecture

software transforms such data into usable information available in tables and graphs. The main table represents the PPEs registry (displaying information relating to the manufacturer, the supplier, the date of purchase, the price, the lead time, the status, the duration of the batteries, the life cycle, the revisions required by law, the dates for replacement of consumables, and the deadlines). A second table records the assignments to operators (registering name, surname, cost center and serial number, in addition to specific usage information such as collection date, possible replacement, delivery, status at the time of collection VS. Status upon delivery, any use in emergency that might have compromised integrity and would require repair/replacement). The application acquires and centralizes peripheral data at the time of each scan and compares such data with the information from the data base. As a next step it returns a simple and direct message to the operator (traffic light logic), where green is used for indicating total usability of the PPE assigned, yellow for partial usability (need for revision/imminent expiry/need to replace consumables/low battery level/..) and red for non-usability (revisions or deadlines not met/damage not repaired/..). Clear system notifications, associated to the light code, explain the problems to the operator. The manager has also the option to add information to the data base, or to access the data base in sole reading mode.

3.10 Pilot Project

The pilot project was divided in different phases. Initially, the setup of the project was carried out by selecting the suitable PPEs, the staff to be involved and the responsible for the delivery of PPEs. The full team selected was trained. The information processes and procedures were defined; the specifications of the selected PPE and safety requirements were analyzed. The calendar of the project was planned. Then, the implementation phase started by applying the NFC tags to the PPEs; at this point it was activated a beta application on the server (to be extended, once tested, to the Cloud) and the project was launched. On the implemented system, numerous measurements of operating times and

errors were conducted in order to evaluate the efficiency, by comparing with previous measurements. Support to the operators was provided along the whole pilot project. The pilot project allowed to gain experience. A new phase of continuous improvement was started, involving hardware, software, processes and procedures.

4 Results and Benefits

This monitoring procedure brings to the company significant benefits, like: a) a greater safety of the Operators; b) time savings compared to the standard monitoring approach; c) the absence of errors in maintenance, which is required to keep efficiency; d) a longer duration of PPEs (e.g. the damage produced by the spillage of the acid of the batteries, compromising the efficiency of the device); e) the timely replacement, once expired, of the consumable materials (like the headphone pavilions), so preventing operators, annoyed by malfunctioning, to reduce the use (a correct management of consumables also allows for a better hygiene for the operators, as in the case of mask filters); f) the reporting procedures introduced allow great transparency on maintenance procedures; g) the register provided records the assignments, reducing the number of losses; h) the list of PPEs available is constantly updated, providing the manager with a real-time image of the equipment to take the most appropriate measures (as urging operators to return PPEs not more needed; or to provide to the replacement of PPEs or spare parts); i) also a greater competitiveness for the company in the tenders, thanks to the differential that it can boast from competition in terms of the safety in the workplace.

5 Conclusions

PPE from life-saving devices can even generate phases of risk if expired or not properly maintained. This paper tackled the issue of a 4.0 management of PPE to keep them efficient during their life cycle and, upon expiry, to alienate them. The intervention is divided into two phases: a) the development of tracking tools for the delivery/use/return, to be compliant with the specifications; b) the conceptualization and development of a device for on-the-job control by the security staff of the critical revision and expiry dates. This device does not interfere with the wearer's activity and can be used both within the company and on external construction sites. The authors would expect that other companies will quickly follow the example of Ansaldo Energia, which felt the need to adopt an innovative methodology to continuously control PPEs efficiency, to safeguard operators' safety. The design architecture studied by the Team has proven, after a severe test phases, to respond to the needs for which it was conceived, for this reason the authors claim that it represents a real asset for Safety 4.0, where it is made use of PPEs. Particular attention was paid by the team to the standardization and generalization of the new procedure, to develop a system that can be used in every operational situation. Considering the small size of the NFC tag and the ease of integration on most of the PPEs on the market, the proposed system can be standardized for extension to other sectors such as, for example, construction, quarries, plumbing and heating. The IIOT nature of the system allows a further generalization to monitor even other types of devices than PPEs (e.g. special tools, equipment,…).

Special thanks to the management, the operators and the labor unions who have made it possible to carry out this activity successfully, in particular to the Eng. Manuelli, Bleve, Santucci, Chiapparoli, Magro and Ziveri and to each member of the project team, as well as to the functions involved, and to the Smart Track and Futuro partners.

References

1. Zhang, L.-S., Li, J., Wang, F., Shi, J.-D., Chen, W., Tao, X.-M.: Flexible stimuli-responsive materials for smart personal protective equipment. Mater. Sci. Eng. R **146**, 100629 (2021)
2. Abrarda, S., Bertrand, M., De Valence, T., Schaupp, T.: French firefighters exposure to Benzo[a]pyrene after simulated structure fires. Int. J. Hyg. Environ. Health **222**, 84–88 (2019)
3. Kelma, A., et al.: Mobile passive Radio Frequency Identification (RFID) portal for automated and rapid control of Personal Protective Equipment (PPE) on construction sites. Autom. Constr. **36**, 38–52 (2013)
4. Shi, J., Li, H., Xu, F., Tao, X.: Materials in advanced design of personal protective equipment: a review. Mater. Today Adv. **12**, 100171 (2021)
5. Farooqui, R.U., Ahmed, S.M., Panthi, K., Azhar, S.: Addressing the Issue of Compliance with Personal Protective Equipment on Construction Worksites: A Workers' Perspective. Corpus ID: 33080169
6. Brouwer*, D.H., Marquart, H., Van Hemmen, J.J.: Proposal for an approach with default values for the protection offered by PPE, under European new or existing substance regulations. Ann. occup. Hyg. **45**(7), 543–553 (2001). © 2001 British Occupational Hygiene Society. PII: S0003-4878(01)00021-7
7. Bernevig-Savaa, M.-A., Darabontb, D.C., Lohana, M., Mihalachea, E., Bejinariu, C.: Selection and verification of personal protective equipment in the context of current legal requirements. Quality-Access Success **20**(S1), 109 (2019)
8. Desjardins-David, I., Arteau, J.: Evaluation of personal protective equipment used for work: considerations and proposed methodology – the criteria to be checked. GfA (Hrsg.), Mensch, Technik, Organisation – Vernetzung (2011)
9. Lee, Y.-A., Salahuddin, M., Gibson-Young, L., Oliver, G.D.: Assessing personal protective equipment needs for healthcare workers. Health Science Reports published by Wiley Periodicals LLC. Health Sci. Rep. **4**, e370 (2021). wileyonlinelibrary.com/journal/hsr2 1 of 9. https://doi.org/10.1002/hsr2.370
10. Ledda, A., Palomba, R.: Innovative wearable systems for improving workers' safety. In: Ahram, T., Taiar, R., Colson, S., Choplin, A. (eds.) IHIET 2019. AISC, vol. 1018, pp. 332–337. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-25629-6_51
11. Sehsah, R., El-Gilany, A.-H.: Personal protective equipment (PPE) use and its relation to accidents among construction workers. Med. Lav. **111**(4), 285–295 (2020). https://doi.org/10.23749/mdl.v111i4.9398
12. Dolez, P.I., Vu-Khanh, T.: Recent developments and needs in materials used for personal protective equipment and their testing. Int. J. Occup. Saf. Ergonomics **15**(4), 347–362 (2009). https://doi.org/10.1080/10803548.2009.11076815
13. Dąbrowska, A., Bartkowiak, G., Kotas, R.: Evaluation of functionality of warning system in smart protective clothing for firefighters. Sensors **21**, 1767 (2021)
14. Campero-Jurado, I., Márquez-Sánchez, S., Quintanar-Gómez, J., Rodríguez, S., Corchado, J.M.: Smart Helmet 5.0 for industrial Internet of Things using artificial intelligence. Sensors **20**, 6241 (2020). https://doi.org/10.3390/s20216241
15. Hashim, M.A.S., May, E.J.: Effectiveness of Personal Protective Equipment (PPE) at Construction Site. eISSN:2600–7920 INTI J. **1**, 2018 (12) (2018)

16. Parvathy, A., Gudivada, V.R.R., Venumadhav Reddy, M., Manikanta Chaitanya, G.: RFID based exam hall maintenance system. In: IJCA Special Issue on "Artificial Intelligence Techniques - Novel Approaches & Practical Applications" AIT (2011)
17. Amicucci, G.L., Settino, M.T., Pera, F.: La manutenzione per la sicurezza sul lavoro e la sicurezza nella manutenzione. © 2019 Inail ISBN 978-88-7484-154-7 (2019)
18. Adgar, A., Addison, D., (Alan) Yau, C.-Y.: Applications of RFID technology in maintenance systems. Corpus ID: 16043737
19. Caponetti, S.: L'obbligazione di sicurezza al tempo di Industry 4.0. DSL 1-2018 (2018)
20. Márquez-Sánchez, S., Campero-Jurado, I., Herrera-Santos, J., Rodríguez, S., Corchado, J.M.: Intelligent platform based on smart PPE for safety in workplaces. Sensors **21**, 4652 (2021)
21. Heidi Lehmann, K.: Smart PPE to prevent heat stress. PSJ Professional Safety, August 2020. assp.org
22. Podgórski, D., Majchrzycka, K., Dąbrowska, A., Gralewicz, G., Okrasa, M.: Towards a conceptual framework of OSH risk management in smart working environments based on smart PPE, ambient intelligence and the Internet of Things technologies. Int. J. Occup. Safety Ergonomics **23**(1), 1–20 (2016)
23. Mosca, R., Mosca, M., Pagano, S., Revetria, R., Galli, G.: Engineering 4.0 to improve the safety of plant operators in a metalworking company of international importance: the Ansaldo Energia Case. In: Proceedings of the IMECS 2021, 20–22 October 2021, Hong Kong (2021)

An Application of Engineering 4.0 to Hospitalized Patients

Roberto Mosca, Marco Mosca(✉), Roberto Revetria, Fabio Currò, and Federico Briatore

Mechanical, Industrial and Transport Engineer Department (D.I.M.E.), University of Genoa, 16126 Genoa, Ge, Italy
mosca@diptem.unige.it, marcotulliomosca@gmail.com, roberto.revetria@unige.it, 4452164@studenti.unige.it

Abstract. In this paper the authors address the problem of surveillance of bedridden patients in hospitals and residences for elderly. Unfortunately, patients cannot be supervised by operators 24 h a day, given the associated costs. An attempt to solve this problem is already provided by wearable devices. This paper describes a 4.0 system implemented to overcome the limits (identified by interviewing a sample of nurses belonging to different facilities) of the wearable devices available on the market. The system proposed consists in monitoring the bed, instead of the patient, through applied sensors. By centralizing and analyzing the data collected it is possible to promptly inform the operative center of the occurrence of risky events to which bedridden patients are normally subjected. The scope of the system is preventing such risks, where possible, or mitigating their effects with a real time intervention. A case study on an active facility, conducted as a pilot project, confirms the humanitarian and economic benefits for patients and facility.

Keywords: Bed management · Bed 4.0 · IOT · Bed · Healthcare 4.0

1 Premise

The paper aims to highlight one of the many applications of Engineering 4.0, in which this discipline can make a positive contribution to the Healthcare sector, without interfering with the "care activity" which is exclusive task of the doctors. Such application is based on a peripheral device designed and created by the authors in both versions:

- Wired (ethernet) or wireless (wi-fi/Bluetooth 5.0) normally offered to hospitals, medical and surgical outpatient clinics, recovery rooms in operating blocks and residences for elderly.
- Telemetry (GSM, GPRS) offered to healthcare facilities for home assistance.

The system allows to monitor and notify the surveillance operators in case of non-conformities, so that they can intervene and anticipate unwanted event. A prompt intervention avoids critical situations as well as to the inevitable economic, criminal and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 235–244, 2023.
https://doi.org/10.1007/978-3-031-16281-7_23

image consequences for the hosting facility. Particular attention in design was paid to reliability and cost. This because, being the device considered by the authors as a tool to alleviate the suffering of bedridden patients, the goal was to ensure the maximum diffusion for the benefit to the largest number of patients achievable.

2 Introduction

The authors have conceptualized and designed a system to fix some serious problems common to hospitals, nursing homes and residences for elderly, with which the development team have been in contact for over 15 years. A major risk identified is the manifestation of painful bedsores bedridden patients for long periods, favored by the droppings and by reduced mobility. Healing from this requires painful and expensive long care. Another problem occurs, in the absence of surveillance, when they try to leave the bed, by climbing over the protective sides. The result can be disastrous, with multiple fractures and hemorrhages, due to falls. Another problem is represented by elderly patients subject to prolonged bed rest, who frequently wander around the structure without reason, with exposure to risk for themselves and for others. When asked about ward doctors and health directors, as highlighted in the literature, they explained that to avoid such events it would be necessary to have 24-h surveillance, which is not compatible with the associated costs. Determined to help fixing a problem that causes further pain to hospitalized people, the authors have resorted to what Engineering 4.0 currently makes available, to create a system capable of perceiving, through appropriate sensors, both the wet bed both the stasis and preparatory movements to climb over the containment barriers. Sudden alerts indicating the bed number and the type of event in place are sent to a central control center. This triggers an operator to intervene. A systematic use of this system will allow to solve this problem, thus reducing pain and high social costs. The phases of conceptualization, engineering and development took over six months. This system, tested in a pilot project (as described in the case study), then extended to more facilities, received enthusiastic comments from the staff, releasing the operators from the responsibility of continuously monitoring the safety of patients. The system allows for more safety for the patients, less responsibility for the staff and less costs for the facilities for causes and compensation.

3 Literature Review

A careful literature review was conducted by researching predominantly on Scopus, WOS and Google Scholar. Over 82.700 papers were found by researching about bed management over these 3 data bases. It was then realized that the keyword “Bed Management” was used for a wide range of applications, even outside healthcare. Therefore, the selection was largely reduced to 28.608 (by filtering with keywords like medicine/nursing/health). Then, restricting again the field by focusing on hospitalization, the number was reduced to 8.856 articles. A last filtration was necessary to focus the actual topic of the study (since bed management in hospitals is mainly referred to the practice of nurses of managing bedridden patients), further reducing the selection to 326 papers, related to the use of technical devices. Such papers were screened, downloaded,

and analyzed. The conclusion of the research was that only 46 articles were related to the topic treated, involving the use of sensors of which 19 are reported in bibliography. This result made the authors aware that literature does not currently treat much about the application of sensors to beds in healthcare. The literature review was then extended to wearable devices, in order to compare the method proposed with possible alternatives. That said, the literature of the sector takes the surveillance of bedridden patients in great consideration and reports that in charged personnel is often insufficient with respect to the demand [14–17]. All articles agree that there are problems with overcrowding of beds and shortages of staff. Adding the aging of the population, the WHO (World Health Organization) predicts that by 2060 the 30% of the European population will be over 65 years of age [16]. In 2019, the article [17] was published, which deals with how it is necessary to use technology to obtain automated solutions for the problems that arise. In 2020 in Healthcare 4.0 [16], is emphasized the "new promising vision for the Healthcare Industry" by the use of Industry 4.0 methodologies. On the same theme the articles [3, 5–8, 10–12]. From literature emerges the tendency of monitoring the patient by using wearable devices located on wrist, chest, lower back, and ankle, for monitoring of parameters to ensure a better supervision and, therefore, a reduction in injuries, especially from falls. It is reported, nevertheless, about limits related to the discomfort of wearing [18, 19].

4 Description of the Device

The system is based on peripheral devices (1 per bed) and sensors (normally 3–4 per device) such as a load cell, a vibration sensor and one for the presence of moisture in the sheets. The weight variation detector senses the patient's attempts to leave the bed when the patient begins to hold on to the safety rails or headboards, while the vibration sensor identifies any pathologies associated with unusual movements and, finally, the humidity detector perceives the presence of liquids. The devices convert these variations into signals which are centralized on a server, analyzed and transmitted to the Operations Center (Fig. 1).

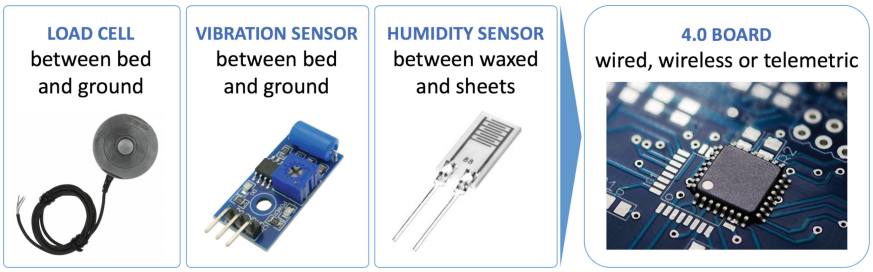


Fig. 1. System peripheral components (bed hardware)

As hardware details, the selected sensors are high compliance and highly efficient components: the load cell guarantees an accuracy of ±1% of span (interval of measure)

with a tolerance of 10 μV/V. The vibration sensor has a capacitance of 750pF with an accuracy ±70. The humidity sensor measure is Boolean (Y/N) with an accuracy of 99,9%. The motherboard developed was successfully tested EMC (Electromagnetic Compatibility) in an anechoic chamber to guarantee not to interfere with any medical equipment of the facility, the test was passed on second attempt after removing any quartz component from the PCB (Printed Circuit Board). The operator immediately requires the intervention of surveillance personnel, thus avoiding such problems. The signal is also recorded in a log file and kept available for statistical processing (Fig. 2).

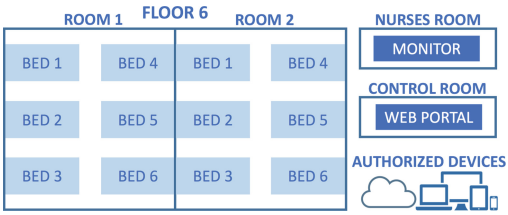


Fig. 2. Layout of the infrastructure to be monitored

In architectural terms (Fig. 3), the use of a concentrator gateway per couple of rooms is envisaged, to integrate the data collected from each bed. The same transmits the data (using different protocols depending on the location, thus avoiding interferences with the medical technologies in use) to a dedicated server, which structures the data in a relational DB, divided by channels, and analyzes them by building information understandable by the operators (as graphs, tables, …). The analysis conducted are consolidated into reports, distributed internally in a hard-wired closed circuit and externally on the Web portal via private cloud, allowing access to authorized devices only, in full compliance with cyber security.

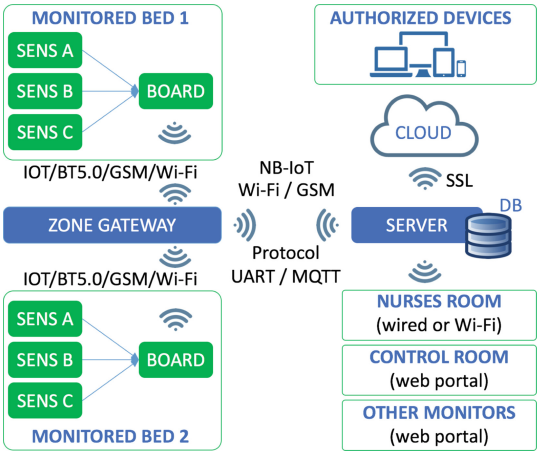


Fig. 3. System architecture

M2M (Machine to Machine) communication, was designed redundant in consideration of the specific needs of installation in different type of environments, therefore it can be selected on the PCB whether to connect via narrow band IOT, Bluetooth 5.0 or even GSM (for centralized surveillance of bedridden patients assisted from home). Therefore, this approach led to the necessity to make available both ethernet wired solution and different non wired alternatives. The flexibility achieved required the use of different protocols of communication, from UART (Universal Asynchronous Receiver-Transmitter) to MQTT (Message Queue Telemetry Transport) and, in order to be compliant to cyber security, the data flow was 24 bit crypted by using SSL (Secure Socket Layer) protocol for any external communications.

5 Other Features

5.1 Stasis Monitoring

The device, by monitoring the center of gravity of the bed, is able to detect if the patient remains immobile for prolonged periods of time (stasis) with the consequent risks for his health. The control center, through the information received from the platform, will real-time assess whether it is a normal situation or it is necessary to take urgent action in order to avoid worsening the patient's condition.

5.2 Tremor Monitoring

In case the patient is prey to abnormal tremors such as those caused by the onset of high fevers or seizures (for patients subject to this type of pathology), the system equipped with a special sensor for monitoring the vibrations of the bed, launches an alert to the central for the necessary measures.

5.3 Weight Tracking

The system is also capable of monitoring the patient's weight. This function allows the structure to measure losses of body mass and increases in patient's weight due, for example, to water retention. Such information allows the medical staff to better assess the patient's condition and to take appropriate measures. It should be noted that the system, by sampling thousands of times per second, is able to distinguish whether the weight variation is real, or it is due to the sudden presence of foreign bodies like books, mobile phones, etc., or exogenous factors, such as drinking a glass of water.

5.4 Standardization

The system has been designed to be compatible with beds from different manufacturers. In this regard, the load cell is installed in a mini platform (3D printed), with safety edges, on which the braked wheel of the bed (or the foot of the bed) surmounts. Another 3 passive platforms, of equal thickness, are used to balance the remaining feet of the bed, to keep the heights aligned.

5.5 Software Features

For a real-time monitoring of the possible states of the patient, a dedicated software was created to allow the central officers to keep the beds under constant control. Figure 4 shows the screen relating to 2 rooms of a given floor. The color code highlights the possible states of the bed-patient system. In case of critical situations, it is enhanced by specific sound signals or even sirens. In this way the surveillance is facilitated with benefit for the patients and for the staff, who is less stressed by the pressure and the responsibility of the work, due to a supporting technology that makes monitoring safe.

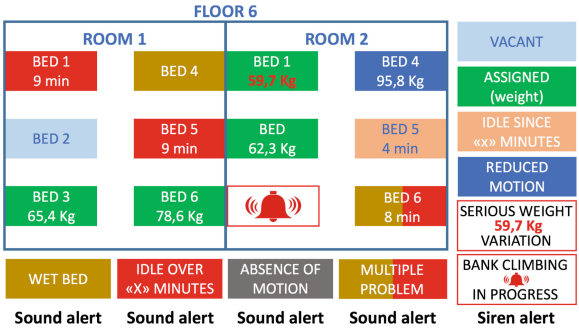


Fig. 4. Monitoring software (layout overview)

Important prerogatives of the system consist in the ability to cover most of the needs of a facility in terms of patient monitoring, allowing the control of each bed in every room. The system may be extended indefinitely, with a plug and play logic. Other types of sensors can be implemented in a second step, in relation to new needs. It is also possible to integrate, because of IOT, any departments and facilities even located at a great distance. The system is easy and intuitive for the operators and the alerts can be set according to the preferences of the same.

6 Innovation

As it emerges from the literature review the application of sensors to beds is still very poorly investigated by researchers, so leaving wide space to innovation. Compared to the use of wearable devices (currently used for a 24 h surveillance of patients) the benefits of the system proposed consist both in releasing the patients, as much as possible, from body applications both in introducing new monitoring functions like droppings identification, fall prevention and weight trend. Contrary to the wearable devices bed management acts in full respect of privacy, since it does not collect any biometric data. Moreover, the data transmission is made via IOT, allowing the centralization of data also in case of multi-site facilities, making information and data accessible via Cloud or Web-portal from any Internet point on the globe, by authorized devices.

7 Benefits

According to the case study presented forward (Sect. 7), the adoption of the proposed system brings a multitude of positive relapses. In fact, it allows the operators to reduce their presence in the departments, thus permitting them to perform other activities. There are also clear benefits for patients in terms of safety and of comfort and for their relatives, who enjoy a higher level of visibility. The facilities who adopt the system, in addition to freeing themselves from responsibilities and costs for the above-described occurrences, will be able to boast and advertise the higher level of service (also in case of domestic assistance) and, therefore, enhance their competitive positioning on the market. In a perspective of continuous improvement, it was included the statistical process control, made easy to understand to any operator through control tables (Fig. 5 shows weight trend). With numerical support, the Management may improve the efficiency from a medical and an economic point of view. In fact, the system allows the facility to decrease costs for causes and compensation due to injuries. As a consequence, it is no wonder the great interest encountered in hospitals, nursing homes, residences for the elderly. The saving achieved translate into an economic advantage, that can be used to improve the service level or to increase the profit for the shareholders.

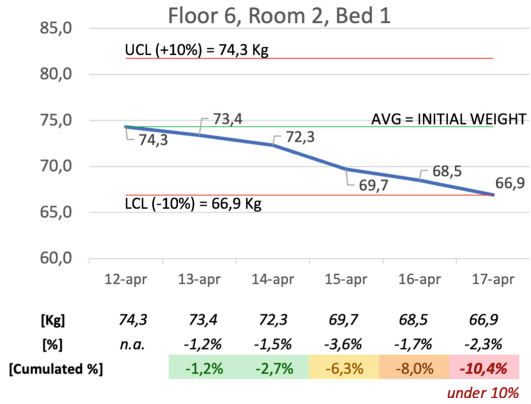


Fig. 5. Monitoring software (patient's weight trend)

8 Future Development

As a new feature the authors believe that the possibility of monitoring the quality of sleep could be of interest, in order to allow doctors to evaluate the opportunity to administer appropriate drugs to give patients an adequate night's rest. In addition, the teams want to investigate the advisability of correlating the effects of certain drugs with any states of impatience/psychomotor agitation of the patient, in order to allow doctors to better calibrate the dosage, or to use substitute drugs, which are better tolerated.

9 Case Study and Economic Sustainability

This paragraph reports the case study relevant to the application of the system described to a nursing home, that hosts full service up to 100 patients aged over 80, with an average presence of 90 patients/day for 365 days/year.

The analysis conducted on historical data highlighted the following:

- Average number of wet beds — 120 / month
- Average number of bed abandoned — 10 / month
- Average number of consequent fractures — 5 / month
- Costs incurred by the institution for repairs — € 6,000 / year
- Costs incurred by the National Health Service — n.a.
- Costs incurred by the institution for legal claims — € 10,000 / year

The data necessary for the economic evaluation are:

- Avoided costs (due to the adoption of the system) — € 15,000 / year
- Costs for the purchase of a system with 100 devices — € 50,000
- Various costs (server, gateway, cables, SW, installation) — included
- Maintenance costs over the life cycle (annual fee) — € 1,000 / year
- Disposal costs — negligible
- System life cycle — 9 years
- Actual savings (€15,000-€1,000) — € 14,000 / year

The PBP (Pay Back Period), calculated as the initial investment divided by the cash flow, is equal to 3.57 years (€ 50,000/€ 14,000). The NPV (Net Present Value), calculated as the sum of the discounted cash flows at the discount rate of 8%, is equal to € 63,989 as shown in Table 1.

Table 1. Financial analysis

	YEAR 0 2021	YEAR	YEAR 10 2031
Discounting exponent	0,0	...	5,0
Expenses			
Investment	50.000 €		
Maintenance cost		...	1.000 €
Income (annual benefit)			
Labor saving		...	15.000 €
Total income	- €	...	15.000 €
Total Expenses	- 50.000 €	...	- 1.000 €
Cash flow (CF)	- 50.000 €	...	14.000 €
Cumulative cash flow	- 50.000 €	...	90.000 €
i = discount rate	8,0%	...	
Discounted cash flow	**- 50.000 €**	...	**9.528 €**
PbP (Pay Back Period)	**3,57**		
NPV (Net Present Value)	**63.989 €**		

10 Conclusions

The basic idea can be defined as an application of Engineering 4.0 to the medical sector. To cope with the problems related to bedridden patients, the authors invested efforts in designing a low-cost system and architecture that would represent a real solution. The outcomes of the study, as certified by the management of the testing facility (Sect. 7), proved that the application of the system brought both the indisputable humanitarian benefits for the hospitalized and the favorable economic implications for the institution.

References

1. Cerchione, R., Centobelli, P., Riccio, E., Abbate, S., Oropallo, E.: Blockchain's coming to hospital to digitalize healthcare services: designing a distributed electronic health record ecosystem. Technovation (in press). ISSN: 0166-4972
2. Cudney, E.A.: A decision support simulation model for bed management in healthcare. Int. J. Health Care Qual. Assur. **32**(2), 499–515 (2019)
3. Ilangakoon, T.S., Weerabahu, S.K., Samaranayake, P., Wickramarachchi, R.: Adoption of Industry 4.0 and lean concepts in hospitals for healthcare operational performance improvement. Int. J. Prod. Perform. Manage. (2021). https://doi.org/10.1108/IJPPM-12-2020-0654
4. Matos, J., Rodrigues, P.P.: Modeling decisions for hospital bed management; a review. In: Proceedings of the International Conference on Health Informatics (HEALTHINF-2011), pp. 504–507 (2011). ISBN: 978-989-8425-34-8
5. Al-Jaroodi, J., Mohamed, N., Abukhousa, E.: Health 4.0: on the way to realizing the healthcare of the future. Received 4 October 2020, Accepted 30 October 2020, Date of Publication 18 November 2020, Date of Current Version 7 December 2020. https://doi.org/10.1109/ACCESS.2020.3038858
6. Sarangi, A., Mohapatra, A., Mishra, T., Keswani, B.: Healthcare 4.0: a voyage of fog computing with IOT, Cloud Computing, Big Data, and Machine Learning. In: Tanwar, S. (ed.) Fog Computing for Healthcare 4.0 Environments. SCT, pp. 177–210. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-46197-3_8
7. Sarangi, A.K., Mohapatra, A.G., Mishra, T.C., Keswani, B.: Healthcare 4.0: a voyage of fog computing with IOT, Cloud Computing, Big Data, and Machine Learning. In: Tanwar, S. (ed.) Fog Computing for Healthcare 4.0 Environments. SCT, pp. 177–210. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-46197-3_8
8. Guidi, G., Pettenati, M.C., Miniati, R., Iadanza, E.: Heart failure analysis dashboard for patient's remote monitoring combining multiple artificial intelligence technologies. In: 34th Annual International Conference of the IEEE EMBS, San Diego, California USA, 28 August–1 September 2012. 978-1-4577-1787-1/12/$26
9. Unterhofer, M., Rauch, E., Matt, D.T.: Hospital 4.0 roadmap: an agile implementation guideline for hospital manager. Int. J. Agile Syst. Manage. **14**(4), 635–656 (2021)
10. Kaur, J., Verma, R., Alharbe, N., Agrawal, A., Khan, R.: Importance of fog computing in Healthcare 4.0. In: Tanwar, Sudeep (ed.) Fog Computing for Healthcare 4.0 Environments. SCT, pp. 79–101. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-46197-3_4
11. Kishor, A., Chakraborty, C.: Artificial intelligence and Internet of Things based Healthcare 4.0 monitoring system. Wireless Pers. Commun. (2021). https://doi.org/10.1007/s11277-021-08708-5

12. Krishnamoorthy, S., Dua, A., Gupta, S.: Role of emerging technologies in future IoT-driven Healthcare 4.0 technologies: a survey, current challenges and future directions. J. Ambient. Intell. Humaniz. Comput. (2021). https://doi.org/10.1007/s12652-021-03302-w
13. Lu, K., Liao, H.: A survey of group decision making methods in Healthcare Industry 4.0: bibliometrics, applications, and directions. Appl. Intell. (2021). https://doi.org/10.1007/s10489-021-02909-y
14. Boaden, R., Proudlove, N.: An exploratory study of bed management. J. Manage. Med. **13**, 234–250 (1999). https://doi.org/10.1108/02689239910292945
15. Howell, E., Bessman, E., Kravet, S., Kolodner, K., Marshall, R., Wright, S.: Active bed management by hospitalists and emergency department throughput. Ann. Intern. Med. **149**, 804–810 (2008)
16. Gardellini, A., Nardi, R., Arienti, V., Panuccio, D., Bernardi, R., Pedone, V.: Aging, patient-bed management and overcrowding in the medical departments. Italian J. Med. **3**(1), 35–43 (2009)
17. Awais, M., et al.: An Internet of Things based Bed-Egress alerting paradigm using wearable sensors in elderly care environment. Sensors **19**, 2498 (2019). https://doi.org/10.3390/s19112498
18. Santos, B., Martins, D., Leao, T., Bock, E.: Supervisory control system for hospital rehabilitation beds. In: 2021 9th International Conference on Control, Mechatronics and Automation (ICCMA)|978-1-6654-1073-1/21/$31.00 ©2021. IEEE (2021)
19. Pang, Z., Yang, G., Khedri, R., Zhang, Y.T.: Convergence of automation technology, biomedical engineering, and health informatics toward the Healthcare 4.0. IEEE Rev. Biomed. Eng. **11**, 249–259 (2018)

A DT-Based System for Predicting Process Behavior

Bowen Qi(✉) and HongSeok Park

University of Ulsan, Ulsan 44610, South Korea
bowen.niceday@gmail.com, phosk@ulsan.ac.kr

Abstract. Manufacturers can use a digital twin to obtain a better understanding of the performance and operational circumstances of a manufacturing asset through near real-time data collected from the asset and make proactive decisions about its optimal operation. Due to its great machining efficiency and good finish quality, grinding is one of the most used precision machining techniques. The vertical double side grinding machine does the necessary work while connected to the PLC program, and we established a methodology for process behavior prediction utilizing a digital twin approach in this paper. The proposed method creates a model of the grinding machine and calculates the grinding force and motor current value using data from sensors, physical models, and system operation. Results from simulations demonstrate how models and communication should be connected. The digital model was established to exactly match the operation of the physical system. Comparison between predicted result obtained from the proposed digital twin model and experiment, revealed a good agreement between proposed model and practice, indicating therefore that the model may be suitable for industrial applications further.

Keywords: Smart manufacturing · Digital twin · Prediction · Grinding

1 Introduction

The competitive manufacturing industry is expanding to smart manufacturing through data-based digitalization to improve production capacity as a result of the emergence of Fourth Industrial Revolution technologies [1, 2]. Considerable attention has been paid to the digital twin (DT), with a focus on improving and growing the global manufacturing industry [3, 4]. To simulate actual objects, DT develops digital virtual models [5]. These virtual replicas can analyze and forecast dynamic changes, transfer optimum values to the actual thing, and mirror the past and current conditions of the physical objects [6]. To allow the optimization of the whole production process, DT may connect the physical and digital worlds in this way [7]. Virtual simulation has the ability to connect real and virtual environments, according to some researchers. A virtual simulation model directly trains on actual assets to extract the right solutions using virtual simulation, accurately reflecting the limits of physical assets. Since it is possible to address issues with physical assets beforehand through trials, it is an efficient instrument for extracting

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 245–251, 2023.
https://doi.org/10.1007/978-3-031-16281-7_24

powerful actions for physical assets and is more advantageous than using the trial-and-error technique on physical assets in terms of time and cost [8, 9].

Double sided grinding is often used for brake discs manufacturing. Due to grinding wheel and workpiece rotated simultaneously in double side grinding process, it's difficult to monitor the grinding force by using dynameter sensor in directly way. For this research work, we developed a methodology for process behavior prediction using a digital twin approach, with the vertical double side grinding machine performing the required work while connecting the PLC program. The proposed approach integrates the information obtained from the sensors, physic models, and operation of the system to establish the grinding machine model and carry out grinding force and motor current value.

2 Conceptual Framework of the System

Since DTRM (Digital Twin Reference Model) has good practicability and scalability, it's providing a general reference model for supporting the application of DT in different fields. Hence, the system framework was designed based on DTRM for brake disc with double side grinding machine in shopfloor. The Fig. 1 illustrate the designed system. There are five model in the system: physical process layer, virtual process layer, DT-based grinding process data, DT-based application service platform, communication, and data mapping mechanism to realize the connectivity among different elements.

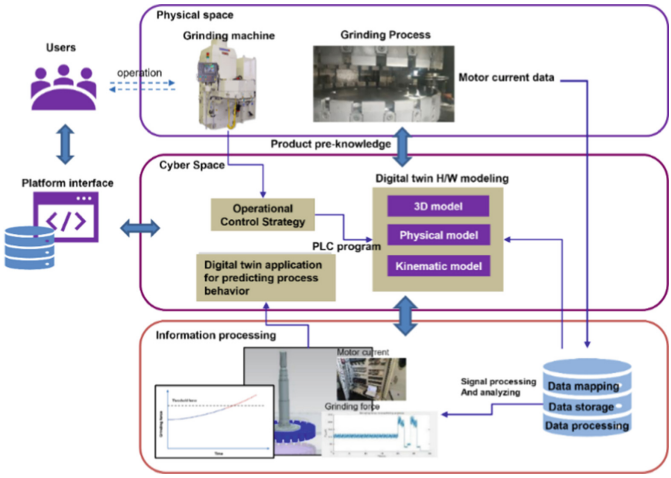


Fig. 1. System framework of proposed digital twin system

3 Development of Digital Twin System

3.1 Modelling of Grinding Machine

The virtual model of the grinding machine contains a 3D geometric model, physical model and kinematic model. The CAD model of the grinding machine was generated

and shown in Fig. 2. The physical model defines the mechanical state of the device in parametric data. The kinematics model defines kinematic pairs of devices for describing the connection between the components of the device.

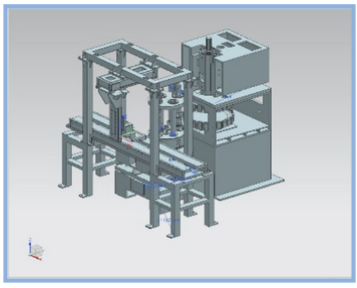


Fig. 2. Modelling of grinding machine

3.2 Modelling of Grinding Process

Due to grinding wheel and workpiece rotated simultaneously in double side grinding process, it's difficult to measure the grinding force by using dynameter sensor in directly way. Therefore, a model which describe the grinding process behavior alone the power flow from motor to grinding wheel was developed. The grinding process behavior can be monitored and predicted by using motor current value which measured in real time.

In order to describe and monitoring real grinding machine behavior, construct a model, which describe the relationship between motor current value and grinding power, by analyzing the mechanical mechanism of components alone the power flow. And then, the grinding process behavior will be also analyzed to establish grinding power model.

The model was programed by using C + + and connect with the virtual process layer. The grinding process behavior can be simulated and described in the real time by input in-process parameters and grinding machine state value.

Power Flow Modelling.
The research object is driven by several motors, this chapter mainly analyzes the upper grinding wheel spindle behavior. All mechanical and electromechanical losses need to analyze to calculate the effective grinding power.

Based on the analysis and modelling, the calculation of effective grinding power algorithm is shown in Fig. 3. This module was programed and connected with database and virtual space layer.

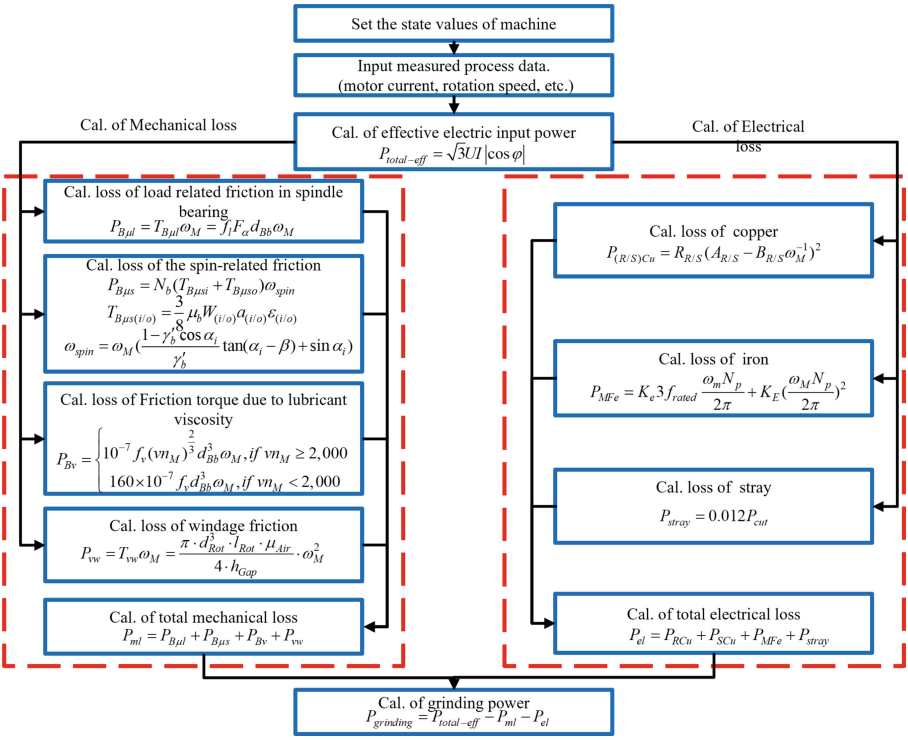


Fig. 3. Calculate effective grinding power

Grinding Process Modelling.

The double side grinding process are shown in Fig. 4. The grinding wheel connected with spindle directly. 24 CBN grinding segment stone evenly distributed in the circumferential direction. The upper and lower grinding wheel, which rotated in different direction, simultaneously process the surface of brake disc. Therefore, the mechanical effect of the upper and lower parts is different and need to be analyzed separately. Firstly, the grinding force of single grinding segment stone will be analyzed. The upper grinding wheel rotated in clockwise and brake disc rotated in counterclockwise. The absolute speed of the grinding stone segment of the upper grinding wheel relative to the workpiece in the Cartesian coordinate system is shown in the following formula.

Then calculate the tangential velocity to the grinding wheel.

$$\begin{aligned} V_x(t) &= \frac{D_1}{2}\omega_1 \sin(\omega_1 t) - \sqrt{l^2 + \left(\frac{D_1}{2}\right)^2 - lD_1\cos(\omega_1 t)}\omega_2 \sin(\omega_2 t) \\ V_y(t) &= \frac{D_1}{2}\omega_1 \cos(\omega_1 t) + \sqrt{l^2 + \left(\frac{D_1}{2}\right)^2 - lD_1\cos(\omega_1 t)}\omega_2 \cos(\omega_2 t) \end{aligned} \tag{1}$$

where**:**

D1: Grinding wheel diameter.

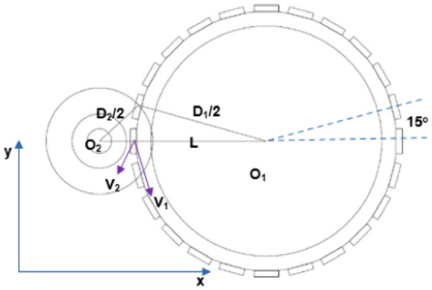


Fig. 4. Grinding process for brake discs

ω1: rotation speed of grinding wheel.
ω2: rotation speed of brake disk.
L: center distance between grinding wheel and brake disk.
The force can be calculated as follow:

$$\begin{aligned} F_{ix}(t) &= K_x \cdot V_{ix}(t) \cdot a_w \cdot a_{ie}(t) \\ F_{iy(t)} &= K_y \cdot V_{iy}(t) \cdot a_w \cdot a_{ie}(t) \end{aligned} \tag{2}$$

where
Kx,Ky,Kz: constants of force coefficients.
V: velocity of grinding stone.
aw: cutting width.
ae: effective cutting depth.

The effective rotation angle of the CBN grinding segment on the workpiece is -18 to + 18. The Fig. 8 illustrate the trajectory of CBN segment relate to rotation angle. The area between the red line is effective area. Calculate the effective tangential force of single grinding stone segment in specific time, and accumulate to get total tangential force:

$$\begin{aligned} F_x(t) &= \sum_{i=n}^{i=m} F_{ix}(t) \\ F_{y(t)} &= \sum_{i=n}^{i=m} F_{iy}(t) \end{aligned} \tag{3}$$

where:
n, m: effective grinding segment number at time *t*.
F_{ix}: *i*th grinding segment force in x direction.
F_{iy}: *i*th grinding segment force in y direction.

Through real-time grinding wheel and workpiece rotation speed monitoring, the effective grinding power is calculated (Fig. 5).

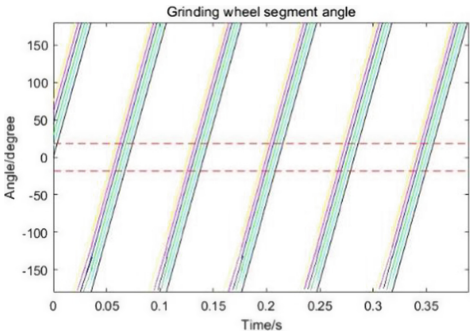


Fig. 5. Effective rotation angle of the CBN grinding segment for grinding

4 Installation of Developed System

At the factory's grinding machine, sensors, DAQ, mountings, and sensor cables are all fitted. DAQ is used to transmit the sensor signals to the monitoring system. The OPC server delivers the real-time process parameters to the monitoring system. Below are the construction statuses for the DATA Integration System (Fig. 6).

(a) (b)

Fig. 6. (a) Connection between physical &virtual space (b) System installation

After the developed digital twin system is correctly connected and installed. By monitoring the real-time processing parameters, including the rotation speed, feed rate motor current, etc., the real-time data process monitoring is shown in the following figure. After the real-time data is obtained in the virtual space, the simulation calculation is carried out, and the predicted spindle motor current value is obtained.

5 Conclusion

This paper demonstrates the methodology to develop and implement a digital twin technical in grinding production line. The system framework was established based on DTRM. In the physical space layer, the physical entities were rearranged based on the design.

Generating the 3D model of grinding production line and give the key corresponding physical parameters to make it have practical significance. The communication module was developed to realize data collection, data processing, and data transfer between physical layer and virtual layer. Due to the characteristics of double side grinding, indirect grinding behavior measurement is proposed and designed in this article. The spindle model was established though analyzing the power losses alone the power flow. Deducing the relationship between the processing parameters, the geometric conditions of the workpiece, grinding force and the grinding power, the grinding process model was generated. The developed model was programmed by using C + + and communicate with virtual space layer to get process parameters, machine state. The grinding process behavior can be simulated and predicted in the real time way.

Comparison between predicted result obtained from the proposed digital twin model and experiment, revealed a good agreement between proposed model and practice. With the help of a digital twin, companies can test and validate a product before it even exists in the real world. By creating a replica of the planned production process, a digital twin enables engineers to identify any process failures before the product goes into production.

In further work in the future, the simulation of the process in the virtual space will be optimized, considering the effects of vibration and heat effect. And develop process control and optimization algorithms and use the established digital twin system to control the actual process to improve the rate of quality products and production efficiency.

References

1. Uhlemann, T.H.J.: The digital twin: demonstrating the potential of real time data acquisition in production systems. Proc. Manuf. **9**, 113–120 (2017)
2. Papetti, A.: A method to improve workers' well-being toward human-centered connected factories. J. Comput. Des. Eng. **7**(5), 630–643 (2020)
3. Rosen, R.: About the importance of autonomy and digital twins for the future of manufacturing. IFAC-PapersOnLine **48**(3), 567–572 (2015)
4. Lim, K.Y.H.: A state-of-the-art survey of digital twin: Techniques, engineering product lifecycle management and business innovation perspectives. J. Intell. Manuf. **31**, 1–25 (2019)
5. Hochhalter, J.D.: Coupling damage-sensing particles to the digital twin concept
6. Qi, Q.: Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. IEEE Access (6), 3585–3593 (2018)
7. Tao, F.: Digital twin-driven product design, manufacturing, and service with big data. Int. J. Adv. Manuf. Technol. **94**(9), 3563–3576 (2018)
8. Swisher, J.R.: A survey of recent advances in discrete input parameter discrete-event simulation optimization. IIE Trans. **36**(6), 591–600 (2004)
9. Barlas, P.: Automation of input data to discrete event simulation for manufacturing: a review. Int. J. Modell. Simulat. Scient. Comput. **7**(1), 163 (2016)

Optimal Robot Workpiece Placement for Maximized Repeatability

Jan Baumgärtner(✉), Philipp Gönnheimer, and Jürgen Fleischer

Wbk Institute of Production Science, Kaiserstrasse 12, 76131 Karlsruhe, Germany
jan.baumgaertner@kit.edu

Abstract. When choosing an industrial robot, repeatability is often one of the most important decision factors. But repeatability is not a global property, instead, it strongly depends on the robot's workspace position. This is currently ignored in robot cell planning leading to a loss of actual repeatability. This paper not only quantifies this loss but presents a new algorithm that finds the optimal position of a workpiece such that repeatability is optimized. The predictions of this algorithm are verified using digital twin simulations on different robots which show up to nine times higher repeatability than with the unoptimized position.

Keywords: Robotics · Robot cell planning · Repeatability · Optimization

1 Introduction

While robots offer much flexibility in today's ever-changing manufacturing landscape, key disadvantages compared to traditional systems are their lower stiffness and repeatability [1]. It is no wonder then that repeatability is one of the most important deciding factors alongside pricing and end-effector velocity [2]. Generally, manufacturers reduce the total repeatability of their robots to a single number, which is calculated according to the ISO 9283 norm. While this is a useful indicator of a robot's overall capabilities, a single number neglects that the actual repeatability of a robot is dependent on the end-effectors' workspace position [3]. By neglecting this position-dependent repeatability during robot cell planning, users potentially reduce the possible repeatability of their robots. This paper presents a new approach capable of increasing the repeatability of a given manufacturing process by simply repositioning the workpiece relative to the robot. The general inputs and outputs of this approach can be seen in Fig. 1.

Fig. 1. Inputs and outputs of the approach presented in this paper

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 252–261, 2023.
https://doi.org/10.1007/978-3-031-16281-7_25

The rest of this paper is divided into four parts: Sect. 2 will give a brief overview of the current state of the art for repeatability estimation and optimization, Sect. 3 will then formulate the mathematical optimization problem, while Sect. 4 will describe the experiments which were used to validate the algorithms as well as their results. The paper closes with a discussion highlighting future research directions in Sect. 5.

2 State of the Art

It has long been known that while helpful, manufacturer-provided repeatability measures are often not as meaningful as one would hope. This is mainly because the calculation depends on self-chosen, often undisclosed, testing points, whose choice influences the final score [2]. This already indicates that the repeatability of a robot manipulator is dependent on its task space position [3]. In the literature, this has been treated as a problem to be investigated experimentally. Prominent examples of such investigations were performed by Kuric et al. [4], Pollák et al. [5] and Mousavi et al. [2]. The reason for this treatment is that repeatability is a highly complex and dynamic phenomenon influenced by several factors such as dynamic flexion or temperature. While many works, such as by Gong et al. [6] have addressed methods to compensate for dynamic influences, the variance of the joint positions remains a lower boundary on the achievable repeatability [6,7]. Works such as by Brethé et al. [3] have thus started modeling repeatability as a phenomenon resulting solely from errors in the joint angles of the robot. Using the model of [3] to predict the repeatability or variance of the end-effector positions, some initial works such as by Kluz et al. [8] have tried to optimize grasp planning. The authors specifically investigated the optimal gripper orientation [8] and x-position [9] for a gripping task. These investigations proved that the repeatability model of [3] can be used to predict and optimize a robot's repeatability. However, the approaches were very limited, being able to optimize only one degree of freedom for one specific robot and only for a single point. The work in this paper vastly extends these previous approaches by optimizing the position and z-orientation of a toolpath for arbitrary robots. This means that given a toolpath and a kinematic model of a robot, the optimizer finds the optimal position for the workpiece such that the robot's repeatability is maximized.

3 Approach

Optimizing the repeatability of a highly nonlinear system such as a robotic manipulator requires a tailored mathematical approach. This section will derive this approach by developing two repeatability measures L adapted from [3] and formulating an optimization problem based on optimal control approaches.

Assuming that the repeatability is solely the product of imprecise joint positioning and not dynamic effects such as temperature changes, the task space

error Δp can be calculated using:

$$p + \Delta p = A(q + \Delta q) \tag{1}$$

where A is the forward kinematics of the robot, p is its task space position and q and Δq are its joint positions and joint position errors respectively. Using a first order approximation of the right-hand side of (1) yields:

$$p + \Delta p = A(q) + \frac{\partial A(q)}{\partial q}\Delta q \tag{2}$$

Leading to the final error approximation of:

$$\Delta p = \frac{\partial A(q)}{\partial q}\Delta q = J(q)\Delta q \tag{3}$$

where $J(q)$ is called the Jacobi matrix. For a simple Scara robot this Jacobi matrix takes the form:

$$J(q_1, q_2, q_3) = \begin{pmatrix} -\sin(q_1) - \sin(q_1 + q_2) & -\sin(q_1 + q_2) & 0 \\ \cos(q_1) + \cos(q_1 + q_2) & \cos(q_1 + q_2) & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4}$$

Note that (3) suggests that the repeatability is dependent on a robot's joint state q and thus only indirectly on its task space position. Since many robots can reach the same position in different joint state configurations this means that the optimal path pose has to be optimized in joint space and not cartesian space. Such optimization requires a scalar measure of the total variance. Based on this formulation of task space error, which was first proposed by Brethé et al. [3] this uniform measure of repeatability $L_{uni}(q)$ can be defined as:

$$L_{uni}(q) = ||\Delta p||^2 = \Delta q^T J(q)^T J(q) \Delta q^T \tag{5}$$

While L_{uni} is very good at describing the overall repeatability of a position, when optimizing one might not be interested in the overall task space error but only specific parts. Consider for example milling. Task space errors along the milling path would not be as detrimental as errors orthogonal to the path (though it should be noted that the proposed algorithm does not consider the significant dynamic forces present during milling). To illustrate this a numerical simulation was set up in which a scara robot was tasked with tracing a planer path. The resulting paths were plotted in Fig. 2. In the left subfigure, one can see that the error along the linear path increases the farther it moves away from $(0, 0)$. This would indicate that the path should instead be kept closer to the body. However, for the singular path in the right subfigure, the deviation from the path seems less when it is farther away. This is not due to the overall error being smaller, but being aligned with the direction of the path. To model this, a second repeatability measure L_{ortho} is introduced which only concerns itself with the repeatability orthogonal to the current path:

$$L_{ortho} = \Delta q^T J(q)^T n n^T J(q) \Delta q^T = ||n^T \Delta p||^2 \tag{6}$$

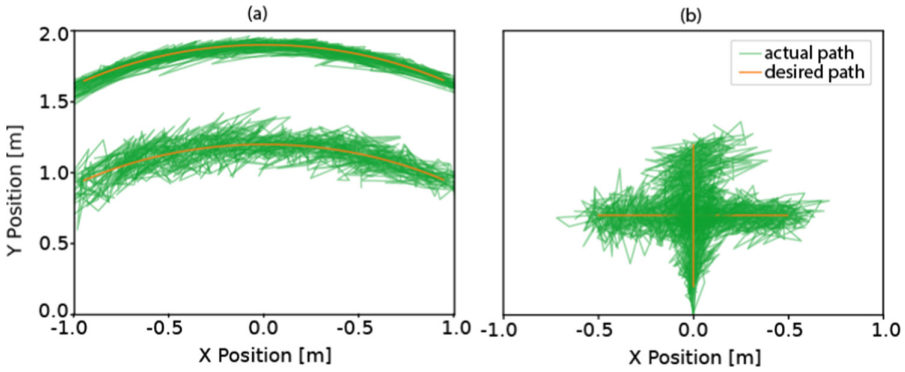


Fig. 2. Numerical path tracing simulation using Scara robot for (a) circular paths and (b) linear paths

where n is the unit vector normal to the path and orthogonal to the z-axis. Using these repeatability measures the optimization problem can be formulated as:

$$\begin{aligned} &\min_{q_1,q_2,\ldots,q_n,b} \sum_{i=1}^{n} L(q_i) \\ &\text{subject to } A(q_i) = s_i(b) \end{aligned} \tag{7}$$

where i denotes the indices of the i-th point of path s and the robot's corresponding end-effector position and joint state. L refers to either one of the two repeatability measures L_{uni} or L_{ortho}. The pose vector b describes how the toolpath s should be moved and rotated to improve repeatability. Since this main objective of the optimization is only expressed through the constraint $A(q_i) = s_i(b)$, the optimization task is ill-posed. Additionally, the joint state does not have to be continuous, leading to large jumps between subsequent joint positions which would be unfeasible in practice. To deal with the first problem a relaxation technique is employed, where instead of forcing the path $s_i(b)$ to be equal to the forward kinematics $A(q_i)$, their distance $||A(q_i) - s_i(b)||^2$ is minimized. The new objective is thus:

$$\min_{q_1,q_2,\ldots,q_n,b} \sum_{i=1}^{n} ||A(q_i) - s_i(b)||^2 + L(q_i) \tag{8}$$

The second problem can be solved by optimizing not over the joint state itself but its relative change $u_i = q_{i+1} - q_i$. This requires a differential kinematic model which can be formulated as follows:

$$\begin{pmatrix} p_{i+1} \\ q_{i+1} \end{pmatrix} = \begin{pmatrix} p_i \\ q_i \end{pmatrix} + \begin{pmatrix} J(q_i) \\ I \end{pmatrix} u_i dt \tag{9}$$

where I is the identity matrix. The optimization problem can thus be formulated as:

$$\begin{aligned} &\min_{u_1,u_2,\ldots,u_n,b} \sum_{i=1}^{n} ||p(q_i) - s_i(b)||^2 + L(q_i) \\ &\text{subject to } \begin{pmatrix} p_{i+1} \\ q_{i+1} \end{pmatrix} = \begin{pmatrix} p_i \\ q_i \end{pmatrix} + \begin{pmatrix} J(q_i) \\ I \end{pmatrix} u_i dt \end{aligned} \tag{10}$$

This new optimization problem bears striking resemblance to classical optimal control-based path tracking problems. These aim to find an optimal set of inputs u to a dynamical system such that the costs along the path are minimized. The main difference is that the problem presented in this paper also allows moving the path itself using the pose vector p. Since this difference does not violate the assumptions of common optimal control approaches, the optimization problem can be solved using optimal control theory. In this case, the multiple shooting approach by [10] was used due to the high nonlinearity of both the problem and the cost function. To improve the numeric stability of the problem the update of q_i and p_i was performed using a 4th order Runge-Kutta method [11] which leads to the following dynamic model:

$$\begin{aligned} x_{i+1} &:= \begin{pmatrix} p_{i+1} \\ q_{i+1} \end{pmatrix} = \begin{pmatrix} p_i \\ q_i \end{pmatrix} + \begin{pmatrix} \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ I \end{pmatrix} u_i dt =: F(x_i, u_i) \\ k_1 &:= J(q_i)u_i \\ k_2 &:= J(q_i + 0.5dtu_i)u_i \\ k_3 &:= k_2 \\ k_4 &:= J(q_i + dtu_i)u_i \end{aligned} \tag{11}$$

where $x_i = (x_{i_1}, x_{i_2}) = (p_i, q_i)$ is the state vector of the dynamics consisting of joint and task space states. To ensure that the end-effector tool is always at the same orientation, the end-effector orientation was constrained to be vertical. The final multiple shooting problem can now be written as:

$$\begin{aligned} &\min_{x_1,x_2,..,x_n,u_1,\ldots,u_n,b} \sum_{i=1}^{n} ||x_{i_1} - s_i(b)||^2 + L(x_{i_2}) \\ &\text{subject to } x_{i+1} = F(x_i, u_i)\ k = 1...n \\ &\qquad O_z(q_i) = e_z \\ &\qquad A(q_1) = p_1 = s_1(b) \\ &\qquad b_{lower} \leq b \leq b_{upper} \end{aligned} \tag{12}$$

Note that this formulation also optimizes the state vectors x_i. This is a trade-off made by the multiple shooting approach where the number of changeable variables is increased while decreasing the nonlinearity of the problem [10]. The first constraint models the system dynamics using the Runge-Kutta update. The second constraint enforces a vertical tool position by forcing the third column $O_z(q_i)$ of the end-effector orientation matrix $O(q_i)$ to be the unit vector along

the z-direction e_z. The third constraint ensures that the initial end-effector position p coincides with the initial joint state q and the start of the path. The constraints b_{lower} and b_{upper} are strictly speaking not necessary but allow a designer to specify where in the workspace the path can be placed. This is helpful since a crowded factory floor often presents space constraints. The multiple shooting problem was solved numerically using gradient-based approaches. Note that this means that only local minima are found. This means that while the optimization will find a pose change b, it might not be the globally optimal one. The state update function (11) was automatically computed from the universal robot description format (URDF) using the trip_kinematics library [12].

4 Experiments and Results

To verify that the optimization improves repeatability, a kinematic Monte Carlo simulation was set up. Here a kinematic model of each robot followed a specified path while noise was added to each joint state. In this case, the noise was modeled as a Gaussian distribution with mean $\mu = 0$ and variance $\sigma = \Delta q$. Unfortunately, the specs of most industrial robots don't contain information about their joint variances. However, for the cost minimization in the solver, the actual absolute magnitude of the variance is less important than their relative size. As an approximation, it is assumed that all joints have the same variance. To showcase potential repeatability improvements, a realistic variance magnitude has to be used. In this case the magnitude was set to $\sigma = 10^{-6}$ rad based on the joint resolution reported in [13]. This is a very conservative lower bound approximation as most joint errors are not a result of the encoder resolution but compliant behavior of the drivetrain or motor shaft twisting [6,13]. It should be noted that the overall magnitude of the variance influences the size of the repeatability measure L relative to the tracking objective $||x_{i_1} - s_i(b)||^2$. As a result, the solver might have trouble converging to an optimal solution. In this case, it can be helpful to multiply L by a scaling factor γ. To verify the solver's results, a sample path had to be defined. Here a figure-eight path was chosen. This general shape of this path can be seen in Fig. 4. The path is parametrized by:

$$\begin{aligned} x(t) &= a\frac{\cos(t)}{1+\sin(t)^2} + x_0 \\ y(t) &= a\frac{\sin(t)\cos(t)}{1+\sin(t)^2} + y_0 \end{aligned} \tag{13}$$

With $t \in [0, 2\pi]$ and a being the scale of the figure and (x_0, y_0) being its center position. Since the solver algorithm was tested using different robots with different workspaces all three parameters were modified for each robot to fit its workspace. While the path parameterization is continuous, the final multiple shooting problem requires a list of discrete positions. Because the number of states to be minimized in equation (12) scales with the number of trajectory points, the number of points influences the complexity of the optimization problem. Since the number of joints of a robot also influences this complexity, the

more complex the robot the fewer points in a path before the solver no longer converges. This means that the maximum number of points in a path is dependent on the complexity of the robot. In this experiment three robots of varying complexity where used: a simple Scara robot a Igus RL-DP robot and a Kuka LWR. The specific degrees of freedom for each robot are shown in Fig. 3. For the first two robots, 20 path points were used, but for the Kuka LWR, only 10 points were used since it has seven degrees of freedom. The visualization of the Monte Carlo simulation for each robot can be seen in Fig. 4. Here the original and for L_{uni} optimized path are pictured. Since the path deviations would normally be too small to see the joint variance was magnified by 10^5. Upon initial inspection, one can see that the optimized path seems to generally have higher repeatability for both the Igus and the Scara robot. For the Kuka however, performance seems equally bad. This is a bit misleading as the plot shows a two-dimensional projection of a three-dimensional problem. It is thus mainly meant to illustrate the difference between the original and the desired path. To evaluate the performance of the solver, a more quantitative analysis was conducted. Here the robot repeated the previous Monte Carlo path following experiment one hundred times while recording the mean absolute path deviation. This position error can be seen in Fig. 5. The quantitative results confirm what Fig. 4 indicated: the

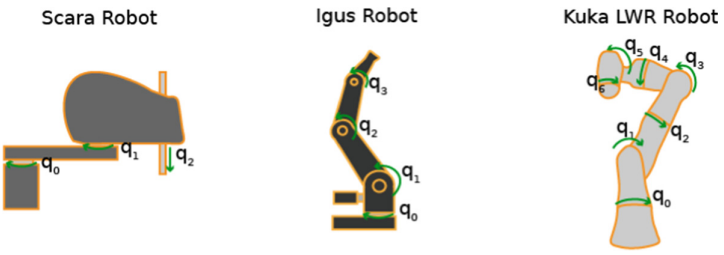


Fig. 3. Robots used for experimental validation

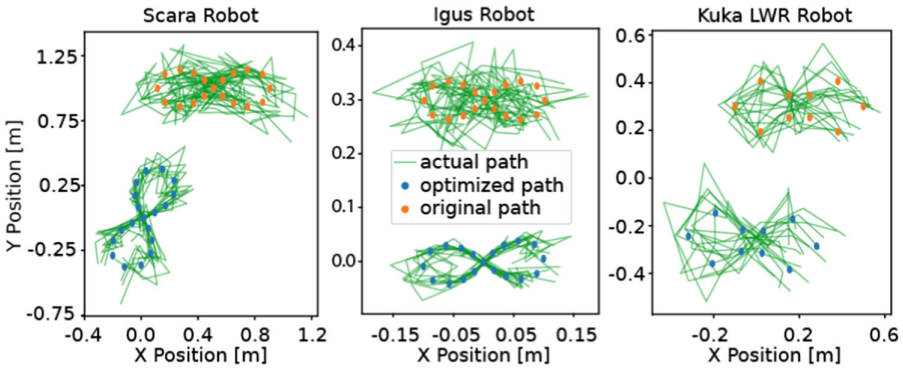


Fig. 4. Repeatability during 20 path repetitions for testrobots with magnified joint variance

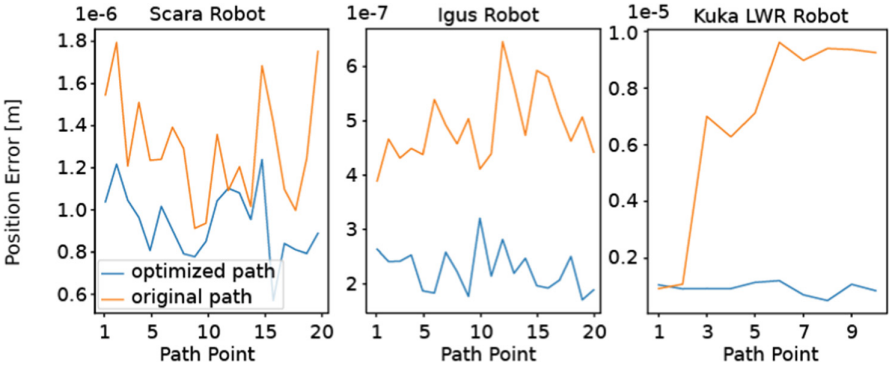


Fig. 5. Mean position error for testrobots

optimized paths have higher repeatability than the original paths. More precisely the optimization provides a maximum relative repeatability gain of 100 % for the Scara robot, 150 % for the Igus robot, and 900 % for the Kuka robot. Interestingly enough, the more degrees of freedom a robot has the higher the relative repeatability gain. This is at odds with the previously explained notion that degrees of freedom add more complexity to the optimization problem. One might have assumed that this leads to more local minimization and therefore worse performance. However, the results indicate that the solver can effectively exploit the additional degrees of freedom for greater performance gains. This could be attributed to the fact that more degrees of freedom lead to more heterogeneous repeatability across the workspace. In the case of the Kuka robot, the solver might also be able to exploit the redundancy in the inverse kinematics. Regardless of the reason, the experiment indicates that repeatability optimization provides more improvements for manipulators with many degrees of freedom. Both experiments above were also repeated for the orthogonal measure L_{ortho}. The visualization of the results can be seen in Fig. 6. Comparing the Figure to the previous result in Fig. 4 one can see that the solver has again found the same optimal path positions. Indeed, the position errors are also identical, which is why they were not shown here. This is contrary to the initial example in Fig. 2 where moving the path outward improved orthogonal errors but not the overall errors. The reason for this discrepancy is likely that the figure-eight structure does not lead itself to the same kind of directional exploitation as a circular path. Though more work has to be carried out to properly investigate the suitability of both measures for different paths and tasks.

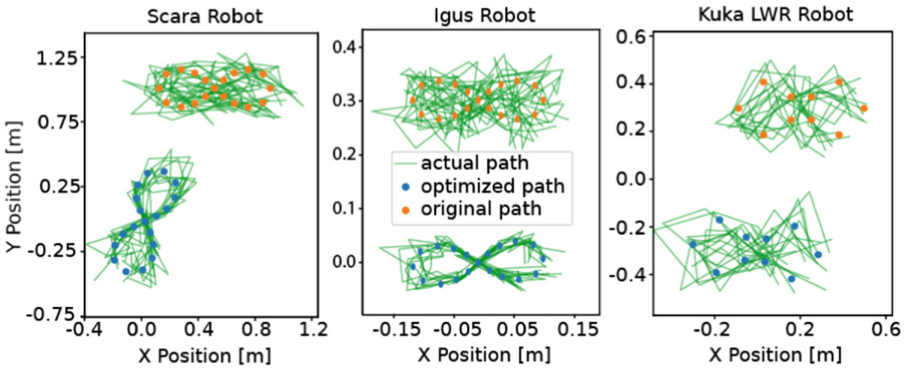


Fig. 6. Repeatability during 20 path repetitions for testrobots with magnified joint variance and optimization based on L_{orth}

5 Discussion

In summary, the experiments showed that the solver is generally capable of optimizing a path. However, as mentioned its convergence is strongly dependent on the number of points in a path. Since most industrial tool paths contain more points than used for these experiments, future work should focus on improving convergence rates for larger numbers of points. The scope of this study was also limited to simulation only. Although prior experiments such as by [8] suggest that optimization of L_{orth} yields improved repeatability in the real world, proper experimental validation of these findings still has to be performed. It should also be noted that the improvements in the real world are bound to be smaller since they also include the aforementioned dynamic influences. There are also several possible extensions of the current algorithm. For example, it forces the tool to always be vertical. This prohibits more complex processes that require varying the end-effector angle from being optimized. Complicated robot welding tasks are a prominent example. To optimize such processes the model of the system has to be extended to also describe the differential kinematics of the end-effector orientation. This is left open for future work. In any case, the results of this paper suggest that optimal control algorithms are uniquely capable of optimizing robot configurations. This opens up an exciting new field of research into the optimization of robotic cells where similar approaches could be used to optimize different values such as energy consumption, stiffness, or even optimal robot choice.

Acknowledgements. The authors would like to thank the Ministry of Science, Research and Arts of the Federal State of Baden-Württemberg for the financial support of the project within the *Innovationscampus Future Mobility.*

References

1. Mühlbeier, E.., Gönnheimer, P.., Hausmann, L.., Fleischer, J..: Value stream kinematics. In: Behrens, Bernd-Arno., Brosius, Alexander, Hintze, Wolfgang, Ihlenfeldt, Steffen, Wulfsberg, Jens Jens (eds.) WGP 2020. LNPE, pp. 409–418. Springer, Heidelberg (2021). https://doi.org/10.1007/978-3-662-62138-7_41
2. Mousavi A., Akbarzadeh A., Shariatee M., Alimardani S.: Repeatability analysis of a SCARA robot with planetary gearbox. In: 3rd RSI International Conference on Robotics and Mechatronics (ICROM), pp. 640–644 (2015). https://doi.org/10.1109/ICRoM.2015.7367858
3. Brethé, J., Vasselin, E., Lefebvre, D., Dakyo, B.: Modelling of repeatability phenomena using the stochastic ellipsoid approach. Robotica **24**(4), 477–490 (2006). https://doi.org/10.1017/S0263574705002481
4. Kuric, I., Tlach, V., Ságová, Z., Císar, M., Gritsuk, I.: Measurement of industrial robot pose repeatability. In MATEC web of conferences, vol. 244, p. 01015, EDP Sciences (2018). https://doi.org/10.1051/matecconf/201824401015
5. Pollák, M., Kočiško, M., Paulišin, D., Baron, P.: Measurement of unidirectional pose accuracy and repeatability of the collaborative robot UR5. Adv. Mech. Eng. (2020). https://doi.org/10.1177/1687814020972893
6. Gong, C., Yuan, J., Ni, J.: Nongeometric error identification and compensation for robotic system by inverse calibration. Int. J. Mach. Tools Manuf. **40**(14), 2119–2137 (2000). https://doi.org/10.1016/S0890-6955(00)00023-7
7. Brandstötter M., Hofbaur M.: The physical upper bound on pose repeatability of general serial manipulators based on joint resolution. In: International Journal of Mechanisms and Robotic Systems, vol. 3, p. 193 (2016). https://doi.org/10.1504/IJMRS.2016.081122
8. Kluz R., Trzepieciński T.: Analysis of the optimal orientation of robot gripper for an improved capability assembly process, Robotics and Autonomous Systems, vol. 74, Part A, pp. 253–266 (2015). https://doi.org/10.1016/j.robot.2015.08.007
9. Kluz, R., Trzepieciński, T.: The repeatability positioning analysis of the industrial robot arm. Assem. Autom. **34**(3), 285–295 (2014). https://doi.org/10.1108/AA-07-2013-070
10. Bock H.G., Plitt K.J.: A multiple shooting algorithm for direct solution of optimal control problems. In: IFAC Proceedings Volumes, vol. 17, Issue 2, pp. 1603–1608 (1984). https://doi.org/10.1016/S1474-6670(17)61205-9
11. Hager, William W.: Runge-Kutta methods in optimal control and the transformed adjoint system. Numer. Math. **87**(2), 247–282 (2000). https://doi.org/10.1007/s002110000178
12. Baumgärtner, J., Millter, T.: TriP: a Python package for the kinematic modeling of serial-parallel hybrid robots. J. Open Source Softw. **7**(71), 3967 (2022). https://doi.org/10.21105/joss.03967
13. Mesmer, P., Neubauer, M., Lechler, A., Verl, A.: Robust design of independent joint control of industrial robots with secondary encoders. Robot. Comput. Integrated Manuf. **73**, 102232 (2022). https://doi.org/10.1016/j.rcim.2021.102232

Enhancing Vendor Managed Inventory with the Application of Blockchain Technology

Santhosh Ganesan, Hendro Wicaksono, and Omid Fatahi Valilai(✉)

School of Business, Social and Decision Sciences, Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany
{Sa.Ganesan,H.Wicaksono,O.FatahiValilai}@Jacobs-University.de

Abstract. As a result of globalization, supply chain networks have grown in complexity and size, posing several issues and chances for development. Inventory optimization and replenishment policy adjustments have a substantial impact on supply chain operating performance and profitability. Vendor Managed Inventory (VMI) is a mutually advantageous agreement between a supplier and a customer in which the supplier oversees inventory and replenishment decisions based on the inventory status of the customer. VMI operations encounter major hurdles in today's supply chains, including trust, data integrity, transparency, and traceability for multiple supplier and customer interactions. Blockchain technology is a distributed ledger that ensures that data is exchanged in a transparent, safe, and secure manner across supply chain stakeholders. The advantages of adopting blockchain technology for VMI operations in a supply chain include decentralized control, security, traceability, and auditable time-stamped transactions. This paper discusses a blockchain-based approach to enhance the VMI supply chain operations. It proposes a generic framework to enable the suppliers and customers order matching in a decentralized mode while fulfilling the trust in terms of managing the data accessibility among the stakeholders as an important prerequisite for VMI establishment. A case study is designed to compare the traditional and blockchain solutions.

Keywords: Supply chain management · Vendor Managed Inventory · Blockchain technology

1 Introduction

Supply chain management is a vital company function that is responsible for planning and controlling all products or services, information, and money flows along the supply chain from the supplier's supplier to the customer's customer. The supply chain comprises people, resources, activities, and organizations that are engaged in satisfying the needs of the customer directly or indirectly by transforming the raw materials into finished products [1]. Many variables enhance the difficulties of modern supply chains, including lack of transparency, disruptions, additional delays, information distortion, and uncertainty. Inventory management acts as an important tool to achieve cost efficiency and balance the demand and supply in the supply chain. In the past, for

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 262–275, 2023.
https://doi.org/10.1007/978-3-031-16281-7_26

inventory planning and replenishment, many strategies, such as demand-driven replenishment, collaborative planning, forecasting, and replenishment (CPFR), Just in Time (JIT), Just in Sequence (JIS), have been implemented. The implementation of new technology increases the supply chain efficiency, resulting in enhanced goods production and shipment. However, many organizations in the industry are unable to cope up with these improvements [2].

VMI is a supply chain optimization method that has been adopted by vendors since it benefits a wide range of stakeholders. Wal-Mart, Hewlett-Packard, and Ericsson Mobile Communication have all reaped significant benefits from adopting VMI arrangements to engage with their suppliers [3]. A VMI arrangement between a supplier and a customer guarantees the supply replenishment independently, with the highest profit margin achievable for both parties. The VMI method is widely used in a variety of industries, including retail, restaurants, hospitality, construction, manufacturing, and healthcare [4]. VMI provides several benefits, including an increase in profit for all supply chain participants when successfully implemented. Such profits are the combined result of cutting down the cost, increasing the sales and related revenues [5, 6]. To establish a VMI relationship between the supplier and the customer, firstly, mutual benefits must be established from both parties. The supplier with access to customer data can predict the demand better for gaining greater control over inventory. On the other hand, the customer can evade the responsibility of planning and processing that results in cost reduction and increases the core competency [7]. VMI's success is built on collaboration and high levels of trust among all stakeholders. The efficient flow of information throughout the supply chain is critical to the effective implementation of VMI [2]. However, there exist several challenges when it comes to establishing a VMI relationship. Some of these challenges include mutual trust between the parties, data integrity, accessibility, and delay in information [8]. Data transparency, data center centralization, and information traceability are among the other issues [9].

Recently, the blockchain technology has been focused by researcher studies with the potential to transform the traditional VMI process [4, 10]. This technology has drawn different scholars and firms to examine it in the field of systematization and supply network management [11]. Furthermore, due to permissioned access to data on this platform, suppliers can more easily retrieve information from numerous data sources such as retailer point of sale, inventories, and deliveries [12, 13]. In addition, with blockchain technology, the planning and implementation processes are made transparent, blockchain synchronizes VMI activities throughout all stages of the supply chain [14].

Considering the trend for the VMI-related literature in the supply chain (SC) domain, there are still several obstacles to its effective implementation especially for enabling the multi supplier and customer support. Furthermore, until now, VMI deployment has been considered through operations management perspectives, whereas concerns of trust, visibility, and security across SC networks have received little attention. This paper tried to fill in the gaps in the concerns described above, as well as extending the concept in terms of process integration where the customer with certain requirements is matched to a suitable supplier who can cater to the needs and derivation of optimal order fulfillment. The proposed solution will enable the multi suppliers and customers VMI interaction

while the issues like data accessibility and trust for the competitive concerns are fulfilled by the blockchain structure.

2 Literature Review

2.1 Vendor Managed Inventory

One of the most important goals of supply chain management is to create a more seamlessly integrated supply chain to improve the performance of an organization [15]. For the effective management of the supply chain, collaboration is the key and acts as an essential element [16]. One such collaborative initiative pioneered by Wal-Mart to increase supply chain efficiency is called Vendor Managed Inventory (VMI). In one of the first VMI initiatives, the company formed such collaboration with Procter & Gamble. VMI is an inventory management process where the buyer or customer relinquishes control of the stocks, and the supplier ensures that sufficient stock is present [17]. To ensure stock availability, the buyer used to assign the minimum and maximum requirements with the implementation of VMI.

VMI helps to reduce the inventory holding cost while maintaining an agreed customer service level. So, it is important to understand and manage the relationship between inventory holding cost and the customer service level [18]. The disclosing of the necessary information such as inventory data, sales data, demand data, promotional activities, etc. to the supplier by the customer help suppliers when it comes to management and aids in the better replenishment decision-making process. The information shared should be precise and on time between the supply chain partners [17].

There are many research papers available that cover the factors affecting the successful implementation of VMI. Some of the important barriers are the willingness of the supply chain partners to share data between them and hesitancy to engage [19, 20], lack of information technology structure, high initial investments [21], the accuracy of shared data between the trading partners [4], ability to make use of the shared information to improve performance [2], geographical distance between the firms [19]. In this Literature review, one of the prerequisites to implement the VMI, trust, is discussed.

2.2 Trust and VMI

Trust is one of the cultural components of relationships with supply chain partners that should be built and maintained by a supply chain-oriented company. Trust comprises integrity and compassion, decide collaboration along with relationship responsibility [22]. Several companies adopted features of a quick response in seeing the success of Wal-Mart, for instance, low inventories, high sales, high customer service level, and stable production cycles, but only a few companies realized its capacity and capability. Companies that were accustomed to the traditional method of safeguarding data to achieve competitiveness resisted modern ideas like trust and collaboration [3]. James and Faizul [23] examined the evolution of relationships between buyer and supplier, proposed a framework for comprehending how buyer-supplier interactions have progressed from arm's length agreements to collaborative procedures built on trust and

sharing of information over the last two decades. They also highlighted the issue of lack of mutual trust. In a VMI relationship, trust exists to be a critical component. For VMI to succeed, both parties must acquire a certain level of trust. Confidential information should be shared with trust; however, there is also a risk associated that the other partner neglects to meet the obligations as decided. Long-term displays of the other partner's performance skills build trust and lower apparent risk [24].

Owing to the absence of mutual trust between the supply chain partners besides lack of adequate information technology, customers share the non-real-time or incorrect inventory and sales data with the supplier. This is one of the two factors that lead to the failure of VMI [25]. Since collaborative initiatives like VMI has a direct positive impact on the supply chain efficiency and indirectly influence the firm's performance, the stakeholders must realize what contributes to more effective collaborative planning decision making. The level of trust between the supply chain partners, as well as the consistency of the data exchanged between the firms, seemed to have an impact on the successful implementation of VMI [26].

2.3 Blockchain and Supply Chain Enhancement

Blockchain is a member of the group of technologies known as distributed ledger technologies [27]. Though the terms blockchain and distributed ledger technology are commonly used interchangeably, the data storage format differs from each other. One can consider blockchain as a subsection of distributed ledger technology, where the blockchain stores transactions in the form of blocks [28]. Blockchains are a new sort of network infrastructure that introduces distributed authenticity, traceability, and agreement via a consensus mechanism to build trust among the stakeholders in networks [29, 30].

In many applications, blockchain enhanced supply chain management. For example, Dwivedi et al. [31] have presented a blockchain-based scheme with a consensus mechanism and smart contracts for securely sharing information in the pharmaceutical supply chain system. Other researchers [27] have discussed an intelligent transportation system with a consensus mechanism for verifying transactions and an authentication protocol for authenticating the user's vehicle. However, there have been only a few articles published in the VMI sector. For instance, a study [32] has discussed the important success factors for adopting a blockchain solution for VMI. It provides a use-case scenario based on a functional smart contract in which a supply chain relationship with a one-to-one relationship between vendor and retailer is considered. As a result, it offers an Ethereum-based blockchain solution that stores data via a decentralized storage system called Interplanetary File System. As an extension, another study [33] has presented a blockchain system with multiple suppliers and customers interacting via four smart contracts. Consequently, it has addressed some of the issues raised in [32]. Finally, a study [14] has offered a proof-of-concept that assesses the feasibility and utility of a VMI-based blockchain system. The proposed solution consists of a functional sequence diagram that depicts the transactional information flow between VMI stakeholders utilizing smart contracts.

2.4 Research Gap Description

Considering the literature review, the VMI requirements for enabling the multi supplier, customer interaction is not approached. This mainly due to trust and data accessibility inside the VMI solution which threaten the competitive advantages of stakeholders. Also, considering the blockchain capabilities, the there is no blockchain-based solution in the literature that seeks to construct an end-to-end, fully functional VMI system. Although these papers have proposed and limited the use of blockchain as a platform to exchange information securely, none have discussed the practical and implementation related aspects of their solutions to enable the suppliers and customers interactions inside the VMI environment. Therefore, this paper aims to provide a system architecture that aids in matching the customers with requirements to a supplier through blockchain and hence providing the holistic view of blockchain-based VMI operations.

3 The Proposed Blockchain Architecture

3.1 Transactions and Block Definitions

The paper uses the proposed architecture of researchers in [10] to set the collaboration workflow. This enables a peer-to-peer distributed cloud platform that permits clients to instantly construct and submit service bundles. Solvers serve as peers in this architecture, can be used to process, manage, and control supplier and customer matching and derivation of order fulfillment policies. Users of this architecture can be visualized as suppliers (service providers) who maintain the inventory of the consumers and the customers (Demanders) who expect their inventory to be maintained within certain limits and never run into an out-of-stock situation. Figure 1 shows the Business process model of proposed blockchain architecture. It is highly essential to encrypt all the messages streaming through this platform since the environment is open to everyone. It allows one to verify the messages and prove the authenticity of their messages. The configuration of the blockchain interaction model will be proposed in proceeding.

3.2 The Proposed Blockchain Structure

Blockchain is made up of an endless number of blocks that are connected chronologically. The block header and body are the two major components of each block in the blockchain. The essential data from the current period's transactions are contained in the body segment. In proposed solution, a transaction mainly focuses on the optimal order fulfillment policy by matching the customer's production requirements and the supplier's capability of meeting the customer's expectations regarding material fulfillment after matching the customer to a suitable supplier. The Public Blockchain Network (PBN) consists of the supplier's capacity, customer's production requirements, and solver pool, which together form a network of nodes on the blockchain. The nodes are updated after each iteration.

Cryptographic keys should be used to ensure the confidentiality of all transactions and fulfill the trust paradigm in VMI. Private/secret keys and public/verification keys are the two types of cryptographic keys needed to accomplish blockchain-based functions

safely. An entity generates private keys, which are then used as a sign to gain access to data. Public keys, on the other hand, are used to provide data access as an authenticator mechanism. These two keys are generated by a requester who needs data access and shares the public key with the data owner while storing the private key. For the secure exchange of information, PBN generates its own private and public keys using the hashing technique. As a result, to index themselves in PBN, each system component, including the customer production requirements pool and supplier's capability pool, must develop its secret hash. With the verification key, every stakeholder in the PBN like customers in (1) and suppliers in (2) can be recognized.

$$CR_{k,public} = Hash\left(CR_{k,private}\right) \tag{1}$$

$$SC_{j,public} = Hash\left(SC_{j,private}\right) \tag{2}$$

when the network has been encrypted, CR_k and SC_j broadcast their requests on PBN and make their verification keys public, allowing all nodes to obtain the data at the same time. Inside the request packets, the required information must be sent, which includes for CR_k – ID, production requirements, location, inventory limits, budget, public key, and so on as in (3); for SC_j - ID, location, price, inventory limits, response time, encryption key, and so on as in (4).

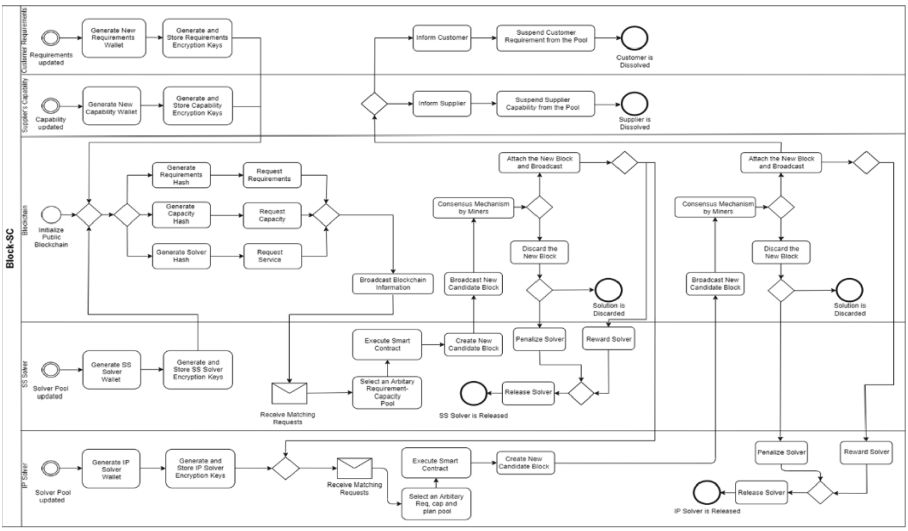


Fig. 1. Business process modelling of adopted architecture for proposed blockchain solution [10].

$$CR_k - Request = Hash\left(ID_k, PR_k, L_k, IL_k, B_k, CR_{k,public} \ldots.\right) \tag{3}$$

$$SC_j - Request = Hash\left(ID_j, L_j, IL_j, P_j, RT_j, SC_{j,public} \ldots.\right) \tag{4}$$

The Supplier Selection (SS) solvers select an arbitrary sub-problem from the supplier's capability and customer production requirement pool after obtaining the necessary data and subsequently generate the smart contract at this point to match the individual customers with suitable suppliers. After the matching problem is solved, candidate blocks are created with each candidate block containing the solution to the supplier selection subproblem. The SS solvers then broadcast the candidate block into the blockchain network, allowing the solution to be verified using the consensus mechanism. The accepted solution of the SS Solvers is fed back to the Inventory Planning (IP) solvers' platform to derive the optimal order fulfillment policy. Once the public key of the IP solvers is shared, access to the corresponding subproblems requested is provided. After the matching requests are received by the IP solvers, the subproblems are selected, the smart contract to derive the optimal order fulfillment policies is executed, the individual candidate blocks containing the solution to the optimal order fulfillment policies are created and broadcasted to every presented node in the system by the IP solvers, the consensus mechanism for every candidate block is executed and at last, the solution is verified.

3.3 Mining Operation

Mining is a critical component of blockchain technology, as it validates the system's transactions. This feature aids the blockchain in ensuring that transactions are recorded correctly and without tampering. A miner is a motivated individual who is authorized to solve the assigned problem and approve incoming transactions. Similarly, block-SC makes use of this function: 'Solvers' are the entities that have replaced the miners' element in the blockchain. Solvers are eager entities; they could be a single person with sufficient processing capacity or a large corporation with powerful servers tasked with resolving the assigned service composition. Solvers, like other PBN entities, must introduce their hash keys to join blockchain as in (5).

$$S_{s,\text{public}} = \text{Hash}\left(S_{s,\text{private}}\right) \tag{5}$$

where S is the solver set.

The SS solvers broadcast requests for appropriate subproblems, which include the solver's ID, capacities, and credibility, as well as its credit score. After the solver selects the subproblem, the blockchain broadcasts an acknowledgment. The associated solver decrypts client production requirements and supplier capability requests using corresponding private keys as soon as the solver receives the acknowledgment. The solver learns about the supply-demand matching issue and the smart contract to be executed after successful decryption. Here, the smart contract can be defined as the best supplier and customer matching. The IP solvers gone through the same process after getting access to the accepted solutions of the SS solvers. Here the smart contract can be defined as optimal order fulfillment policy. When a solver solves a subproblem, it creates a new candidate block that contains all necessary information about the problem and solution. The solution consists of the suppliers and customers cluster in which the suppliers would supply the customer requirements. The cluster is proposed a block node to the blockchain platform. This block is broadcasted to the entire system. To check the legitimacy of this

block, a consensus mechanism would be triggered right after its submission. It is added to the PBN if it complies with the consensus mechanism and is found to be legitimate and untampered. Generally, it is expected that the number of volunteer solvers in the system is projected to exceed the number of sub-problems, some may be assigned multiple times. If a solver's subproblems have any domain overlap with the winner solver, it must drop/modify all of its calculations and start a new process as soon as the new candidate block verifies; otherwise, it can continue with its present calculation.

3.4 Consensus and Reward Mechanism

A Proof of Work (PoW) consensus mechanism in proposed architecture is consider as:

$$\text{Block}_x = \text{Hash (CP, PR)} \tag{6}$$

where CP and PR are the cryptogenic puzzle and proof.

Any change in the incoming message causes the hash string to be completely different. As a result, any changes in the customer production requirements, as well as the supplier's capability, generate a new hash string for the candidate block. As a result, if an incoming message has been tampered with, the current hash and previous hash do not match, the message cannot be added to the blockchain. To incentivize the solver to solve the (sub) problems, blockSC incorporates a reward structure adapted from cryptocurrencies.

3.5 Executive Procedure

The business processes and interactions will be managed by updating the requirements of the customer and capabilities of the supplier in the architecture. Later, these are stored as a wallet. Then the generation of the private and public keys for the respective wallet takes place.

This architecture uses two levels of solvers – Supplier Selection solver (SS solver) and Inventory Planning solver (IP solver). At first, only the SS solvers will take part in the architecture to match the customer with a suitable supplier. Once this process is accomplished, the IP solvers derive the optimal order fulfillment policy i.e., at which time, which amount of goods has to be made available. The SS solvers enter the architecture by generating and storing the SS Solver wallet and their respective encryption keys. The PBN that serves as the intermediate platform between the data owners and data requesters generates its own private and public keys. Then the supplier with certain capabilities and the customer with certain requirements broadcast their information in the PBN. Subsequently, the data is hashed via an encryption mechanism.

Thereafter, Solvers must select the appropriate subproblems, in this case, supplier selection, based on their solving capabilities. The PBN and the SS solvers share their public keys to each of the corresponding subproblems in the supply and demand pools. The required information needs to be sent inside the request packages and autonomous access to the respective subproblems is provided to the SS solvers. SS solvers sequentially select the subproblems and execute the smart contract to match the individual customers with the suitable supplier. Several candidate blocks are created with each candidate block

containing the solution to the supplier selection subproblem. Right after the creation of a new block by a solver, the new block would be broadcasted to every presented node in the system. Subsequently, the consensus mechanism is executed for every candidate block and the solution is verified.

The accepted solutions are broadcasted and stored in the PBN. On the other hand, the rejected solutions are discarded from the architecture. It is noteworthy that the SS solvers are rewarded for the accepted solution. Whereas the SS solvers are penalized for the discarded solutions. As soon as, the reward mechanism is executed, the SS solvers leave the architecture. Next, the IP solvers enter the Block-SC architecture by generating and storing the IP solver wallet and their respective encryption keys. Thereafter, the accepted solution of the SS Solvers is fed back to the IP solvers' platform to derive the optimal order fulfillment policy. Once the public key of the IP solvers is shared, access to the corresponding subproblems requested is provided.

After the matching requests are received by the IP solvers, the subproblems are selected, the smart contract to derive the optimal order fulfillment policies is executed, the individual candidate blocks containing the solution to the optimal order fulfillment policies are created and broadcasted to every presented node in the system by the IP solvers, the consensus mechanism for every candidate block is executed and at last, the solution is verified.

Once the accepted solutions are broadcasted, they are stored in the PBN, and the suppliers and the customers are informed about the corresponding pairs and order fulfillment policies. The fact that the SS solvers are rewarded for the accepted solution is noteworthy. The SS solvers, on the other hand, are penalized for abandoned solutions. Hence, the requirements of the suppliers, customers, and IP solvers are fulfilled, they are dissolved from the system. This completes the architecture.

4 Case Study Scenario

The following scenarios explain the supplier and customer matching and VMI based order fulfillment derivation facilitated in the architecture. In the scenarios, three suppliers form the supply pool with certain capabilities and three customers form the demand pool with certain requirements have been considered. Without blockchain application the implementation of VMI among the three suppliers and three customers would not be possible in a centralized model. This is due to accessibility of data among all suppliers and customers which threaten their competitive advantage. So, using the proposed model in this paper, the blockchain platform can insure the data regarding the transactions and capabilities of suppliers and requirements of customers are maintained both integrated in the platform and also would be only accessible regarding assignment policies which SS solvers would define. Further, the SS solvers match the customer from the demand pool to a suitable supplier from the supply pool based on the cost i.e., price expectation of the customer and price offered by the supplier and quantity. Table 1 shows the supply pool and demand pool with its attributes.

Table 1. Supply and demand pool

Supply pool			
Attributes	Supplier 1	Supplier 2	Supplier 3
Location	Bremen	Hamburg	Erlangen
Product	Bolt	Nut	Bolt
Quantity Ava. 1 Dec 2021	1000	2500	2200
Price 1 Dec 2021	2.1$	3.05$	1.98$
Quantity Ava. 1 Jan 2022	1500	2500	1800
Price 1 Jan 2022	2.02 $	3.02$	2$
Lead time (Days)	3	2	4
Demand pool			
Attributes	Customer 1	Customer 2	Customer 3
Location	Kassel	Berlin	Hamburg
Product	Toy Car	Toy Car	Table
Raw material requirement	Boltx1, Nutx1	Boltx2, Nutx2	Boltx4, Nutx4
Current inventory level	800;800	500;1400	800;600
Production Plan 5 Dec 2021	600	350	200
Production Plan 7 Jan 2022	500	450	500
Max & Min inventory limits	950;900	2000;1500	900;1000
Price expectations	2.1$;3.10$	2$;3.05$	1.96$;3.07$

The objective of the IP solvers is to derive the optimal order fulfillment policy. In this illustration, the IP solvers employ a constant demand model. The advantages of the constant demand model are multifarious – steady inventory consumption, production leveling & smoothing, and high-capacity utilization. The decision variable for the prescribed constant demand model is schedule receipts (S_t), which are changing day to day. The constraints are the maximum and minimum inventory levels. In VMI, the suppliers must maintain the inventory levels which are agreed mutually before establishing a VMI relationship. Otherwise, the suppliers are entitled to pay penalties. Therefore, maintaining the inventory on hand (I_t) within the maximum and minimum limits is crucial. Thus, in aiding the suppliers with deriving the optimal order fulfillment policy, IP solvers are employed in this architecture. Consequently, based on the accepted solution of the SS Solvers, IP solvers derive the optimal order fulfillment policy using the demand netting method which transforms gross requirements (Gt) into net requirements (Nt), which in turn indicates the time and amount of goods which should be made available. Gross requirements (G_t) can be derived by Gt = TD/Nw and St can be calculated by $I_{max} - I_t + G_t$, when $I_t < I_{min}$ and 0 when $I_{min} < I_t < I_{max}$ while $I_t = I_{t-1} - G_t + S_t$ Subject to $I_{min} \leq I_t \leq I_{max}$ and $N_t = \min(\max(-I_t,0), G_t)$.

4.1 Sample Scenario of Solver Solution Overlap

Table 2 shows the possible solution for a sample month proposed by the SS solvers. As assumed each supplier can be mapped to more than one customer i.e., multiple sourcing. Solver 1 develops order fulfillment by supplier 1 for the customer 1 while Solver 3 develops order fulfillment by supplier 2 and 3 for the customer 1 and 2 in parallel. As soon as the Solver 1 assign the responsibility for supplier 1 and 3 through its proposed solution, the validation process would be initiated in the blockchain platform. Assuming that the provided solution is valid, it would be attached to blockchain as shown in Table 3. Considering that solver 1 and solver 3 are fulfilling the requirements of the same customer 1 with the capabilities of different suppliers, as the solution of solver 1 is attached to the blockchain, solver 3 has to make a choice: (1) drop its progress and restart for a new sub-problem solving and proposal to the blockchain; (2) continue with the current sub-problem but it has to purge the suppliers and enhances the scope of the problem in terms of fulfilling other customers. If the solver3 choose option 1, it has to face with the fact that all of its computation progress was in vain, and this action would be a pure loss as it would receive no reward. If it chooses the second option, it may have the chance to use part of its computation progress and propose an enhanced solution to the evolved scope of the problem like fulfilling another customer requirement by supplier 2 and 3 capabilities. In this case, a few extra steps would be imposed on the solver, but it can then broadcast its block as a new assignment to blockchain and gain reward.

Table 2. SS solvers solution proposal coverage (+ solver1, * solver 3)

			Customer 1	Customer 2	Customer 3
		Bolt	$2.10	$2.00	$1.96
		Nut	$3.10	$3.05	$3.07
Supplier 1	Bolt	$2.02	+		
Supplier 2	Nut	$3.02	*	*	
Supplier 3	Bolt	$2.00	*	*	

As described in the sample scenario, the blockchain platform is enabling the multi suppliers and customers assignments with the SS solvers. As the blockchain is equipped with the smart contract mechanisms and the proposed order assignments are encrypted with private and public keys of suppliers and solvers, all the information regarding the requirements of customers and capabilities of the suppliers are preserved from unauthorized access. So only, the SS solver who is proposing order fulfilments among their selected suppliers and customers would have access to the details with smart contract structure. This will enable the data security and also trust among all the stakeholders inside the blockchain platform.

Table 3. Optimal order fulfillment policy proposed by solver 1 for suppliers 1 and 1

Supplier 1: Material – Bolt for Customer 1				
No. of working days: 20				
I_t: 800; I_{max}: 950; I_{min}: 900				
Days	Gross requirements (G_t)	Scheduled receipts (S_t)	Projected on-hand (I_t)	Net requirements (N_t)
1	25	175	950	0
2	25	0	925	0
3	25	0	900	0
4	25	75	950	0
5	25	0	925	0
6	–	–	–	–
7	–	–	–	–
8	25	0	900	0
9	25	75	950	0
10	25	0	925	0

5 Conclusion

The illustrated architecture acts as a bridge between suppliers, customers, and solvers. In this architecture, the process takes place sequentially. Here the consensus mechanism is executed two times, the first time after the supplier and customer matching problem and again after the derivation of optimal order fulfillment policy. The solvers after solving their problem get a reward. Therefore, in this architecture, the reward-giving process also takes place two times, after the supplier selection and the inventory planning business processes respectively. In the modern-day business environment, collaboration is the key to remaining competitive. Moreover, the integration of third parties in the business process is essential for business success. This architecture offers such a collaborative environment to its clients. To establish a VMI relationship between suppliers and customers, trust is one of the important prerequisites. Therefore, the suppliers and customers who are willing to enter into a VMI relationship can effectively utilize this architecture to eliminate the trust factor as a prerequisite. Furthermore, the architecture is propelled by blockchain, which ensures security, transparency, and traceability. Further investigation should be conducted on how to create a win-win situation in the architecture that emphasizes global optimum by incorporating additional constraints in the architecture which ensures equal benefits to all the clients, ultimately proliferating its usability.

References

1. Mentzer, J.T., et al.: Defining supply chain management. J. Bus. **22**, 1–25 (2001)
2. Omar, I.A., Jayaraman, R., Salah, K., Debe, M., Omar, M.: Enhancing vendor managed inventory supply chain operations using blockchain smart contracts. IEEE Access **8**, 182704–182719 (2020)
3. Ireland, R., Bruce, R.: CPFR - Only the beginning of collaboration. Supply Chain Manag. Rev. **SEPT/OCT**, 80–88 (2000)
4. Ahmadi, E., et al.: Using blockchain technology to extend the vendor managed inventory for sustainability. In: International Journal of Industrial and Manufacturing Engineering, ICSCLE 2021: International Conference on Supply Chain and Logistics Engineering, vol. 15, pp. 470–474 (2021)
5. Fatahi Valilai, O., Houshmand, M.: A platform for optimisation in distributed manufacturing enterprises based on cloud manufacturing paradigm. Int. J. Comput. Integr. Manuf. **27**, 1031–1054 (2014)
6. Aghamohammadzadeh, E., Malek, M., Valilai, O.F.: A novel model for optimisation of logistics and manufacturing operation service composition in Cloud manufacturing system focusing on cloud-entropy. Int. J. Prod. Res. **58**, 1987–2015 (2020)
7. Disney, S.M., Towill, D.R.: The effect of vendor managed inventory (VMI) dynamics on the Bullwhip Effect in supply chains. Int. J. Prod. Econ. **85**, 199–215 (2003)
8. Lee, H.L., Whang, S.: Information sharing in a supply chain. Int. J. Manuf. Technol. Manag. **1**, 79–93 (2000)
9. Radmanesh, S.-A., Haji, A., Fatahi Valilai, O.: Blockchain-based cloud manufacturing platforms: a novel idea for service composition in XaaS paradigm. PeerJ Comput. Sci. **7**, e743 (2021)
10. Aghamohammadzadeh, E., Fatahi Valilai, O.: A novel cloud manufacturing service composition platform enabled by Blockchain technology. Int. J. Prod. Res. **58**, 5280–5298 (2020)
11. Gonczol, P., Katsikouli, P., Herskind, L., Dragoni, N.: Blockchain implementations and use cases for supply chains-a survey. IEEE Access **8**, 11856–11871 (2020)
12. Navendan, K., Wicaksono, H., Fatahi Valilai, O.: Enhancement of crowd logistics model in an E-commerce scenario using blockchain-based decentralized application BT - dynamics in logistics. In: Freitag, M., Kinra, A., Kotzab, H., Megow, N. (eds.) Dynamics in Logistics. LDIC 2022, pp. 26–37. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-05359-7_3
13. Khaturia, R., Wicaksono, H., Fatahi Valilai, O.: SRP: a sustainable dynamic ridesharing platform utilizing blockchain technology BT - dynamics in logistics. In: Freitag, M., Kinra, A., Kotzab, H., Megow, N. (eds.) Dynamics in Logistics. LDIC 2022, pp. 301–313. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-05359-7_24
14. Kolb, J., Hornung, J., Kraft, F., Winkelmann, A.: Industrial application of blockchain technology – erasing the weaknesses of vendor managed inventory (2018)
15. Lambert, D.M., Cooper, M.C., Pagh, J.D.: Supply chain management: implementation issues and research opportunities. Int. J. Logist. Manag. **9**, 1–20 (1998)
16. Horvath, L.: Collaboration: the key to value creation in supply chain management. Supply Chain Manag. An Int. J. **6**, 205–207 (2001)
17. Hammer, H., Bernasconi, C.: Best Practice in Implementing VMI A recommendation by ECR Europe (2016)
18. Barratt, M.: Understanding the meaning of collaboration in the supply chain. Supply Chain Manag. **9**, 30–42 (2004)
19. De Toni, A.F., Zamolo, E.: From a traditional replenishment system to vendor-managed inventory: a case study from the household electrical appliances sector. Int. J. Prod. Econ. **96**, 63–79 (2005)

20. Kulp, S.C.: The effect of information precision and information reliability on manufacturer-retailer relationships. Account. Rev. **77**, 653–677 (2002)
21. Mcbeath, B.: The truth about VMI - revelations and recommendations from the 2003 ESCA-chain link research study on VMI in the high tech supply chain. ESCA Chain. Res. (2003)
22. Min, S., Mentzer, J.T.: Developing and measuring supply chain management concepts. J. Bus. Logist. **25**, 63–99 (2004)
23. James, H., Faizul, H.: From arms-length to collaborative relationships in the supply chain. Int. J. Phys. Distrib. Logist. Manag. **30**, 750 (2000)
24. Vigtil, A.: Information exchange in vendor managed inventory. Int. J. Phys. Distrib. Logist. Manag. **37**, 131–147 (2007)
25. Sari, K.: Exploring the benefits of vendor managed inventory. Int. J. Phys. Distrib. Logist. Manag. **37**, 529–545 (2007)
26. Petersen, K.J., Ragatz, G.L., Monczka, R.M.: An examination of collaborative planning effectiveness and supply chain performance. J. Supply Chain Manag. **41**, 14–25 (2005)
27. Dwivedi, S.K., Amin, R., Vollala, S.: Blockchain based secured information sharing protocol in supply chain management system with key distribution mechanism. J. Inf. Secur. Appl. **54**, 102554 (2020)
28. Guggenberger, T., Schweizer, A., Urbach, N.: Improving interorganizational information sharing for vendor managed inventory: toward a decentralized information hub using blockchain technology. IEEE Trans. Eng. Manag. **67**, 1074–1085 (2020)
29. Kumar, A., Mahindru, T., Shukla, P., Sharan, A.: Blockchain: The India Strategy (2020)
30. Cresitello-dittmar, B.: Application of the blockchain for authentication and verification of identity. Int. J. Adv. Sci. Eng. Inf. Technol. **6**, 1–9 (2016)
31. Dwivedi, S.K., Amin, R., Vollala, S., Chaudhry, R.: Blockchain-based secured event-information sharing protocol in internet of vehicles for smart cities. Comput. Electr. Eng. **86**, 106719 (2020)
32. Casino, F., Dasaklis, T.: Improving vendor-managed inventory strategy based on Internet of Things (IoT) applications and blockchain technology (2019). https://doi.org/10.1109/BLOC.2019.8751478
33. Casino, F., Dasaklis, T., Patsakis, C.: Enhanced vendor-managed inventory through blockchain (2019). https://doi.org/10.1109/SEEDA-CECNSM.2019.8908481

Design and Development of a Tomato Picking Soft Robotic Gripper with a Separator and Mechanical Iris Based Pedicel Cutting Mechanism

Shahid Ansari(✉) and Bishakh Bhattacharya

Indian Institute of Technology Kanpur, Kanpur 208016, India
shahid@iitk.ac.in

Abstract. Picking tomatoes from plants in farm fields using an autonomous robotic gripper is a challenging problem especially when tomatoes are present in the form of a dense cluster as multiple aspects have to be kept in mind while doing the successful operation to prevent them from any kind of damage. Here we have proposed a novel compliant and soft robotic gripper with fingers in the form of curved beams made up of thermoplastic polyurethane (TPU) with a separating mechanism for picking target tomato from a cluster attached with a pedicel cutting mechanical iris mechanism having sharp blades. To actuate the gripper fingers a servo motor pulls them through a set of fishing lines. A comparison is drawn for the variation of fingertip positions, velocities, angle w.r.t time, and trajectory followed obtained through kinematic equations and motion tracking through a camera. Finally, a comparative analysis is done for the variation of separating, cutting, grasping, and total operation time w.r.t tomato sample diameter.

Keywords: Soft gripper · Tomato picking · Cutting mechanism · Cable actuation · Servo motor

1 Introduction

In precision agriculture, intelligent robots are rapidly becoming popular due to the multiple advantages such as accurate crop health inspection, precise pesticide spraying, pruning, and harvesting using advanced artificial intelligence-based smart autonomous system designs. The majority of the complex picking and manipulating tasks of the crop is done by the robotic manipulator using a crucial component which is its end effector which can act as a gripper. The gripper end-effector plays a crucial role in handling the products such as fruits, vegetables, and other farm items especially performing the picking and sorting operations. Due to the delicate nature of these products they can get easily bruised or crushed if the handling is improper. Hence, compliant grippers which can be soft [1] yet strong enough [2] are required to perform desired operations on it. Various parameters such as shape, texture, and size of the product are

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 276–286, 2023.
https://doi.org/10.1007/978-3-031-16281-7_27

necessary to control these soft grippers and for secure grasp, an accurate understanding of slip, soft contact, and force control is required [3–5]. Our concern here is mainly focused on tomato picking from the farm fields which is one of the easily damageable products if not handled properly. Some researchers [6] developed the relationship among various parameters such as plastic deformation of food product during grasping, finger operating velocity, and gripper stiffness during the operation. They proposed mathematical functions governing the food behavior through which an optimum set of the design parameters for the robotic gripper to achieve the minimum plastic deformation of food during two-finger gripper grasping operation can be obtained. For robotic harvesting of delicate and soft crops researchers proposed different types of gripper applications ranging from tomatoes [7], oranges [8], sweet pepper [9], strawberries [10–12] to cucumber [13]. The successful picking of the tomato from the plants depends upon three operations which are: (a) separating the target tomato from the dense cluster of the tomatoes (b) pedicel cutting [11,14] and (c) firmly grasping without damage. Various grippers have been developed so far for fruit and vegetable picking such as the use of suction cups and cutting the pedicel by using the sharp cutting edge of the plier like cutter operated with the help of a cable at its handle end for tomato picking [15]. For sweet pepper harvesting [16,17] proposed a thermal cutting system robot based on current and voltage potentials. For tomato cluster harvesting a robot with an end-effector having two fingers with a tomato pedicel cutting mechanism is also reported [18]. Picking of tomato using robotic arm installed with geared scissor cutting mechanism was proposed in [19]

The objective of this study is to design and develop a novel soft robotic gripper for tomato picking having the capability of separating the tomato from a dense cluster and detaching it from its pedicel using an iris [20] pedicel cutting mechanism with curved blades made up of mild steel. The separator and cutter were installed at the entry face of the gripper in addition the soft compliant fingers firmly grasp the subject in its place for operating in a natural horticulture environment in a step-by-step manner for the successful operation.

Compared to tomato picking systems discussed in [15] and [18] this system is different as it is designed to pick the tomatoes from the cluster by performing three operations i.e. separating the target tomato from the other tomatoes, cutting its pedicel and finally grasping, while none of the other systems addressed above used this strategy rather they used to pick up either the whole cluster using a cutter and grasper or used a suction pad to pick the target tomato from cluster and grasp with two fingers which does not ensure its picking without causing any kind of damage.

2 Design

Picking vegetables like tomatoes from the crops in farm fields requires skills to prevent any kind of damage (crushing or bruising) while pruning especially if they are present in the form of dense clusters. If this task is performed using a rigid robotic gripper then it needs precise control to prevent it from any kind

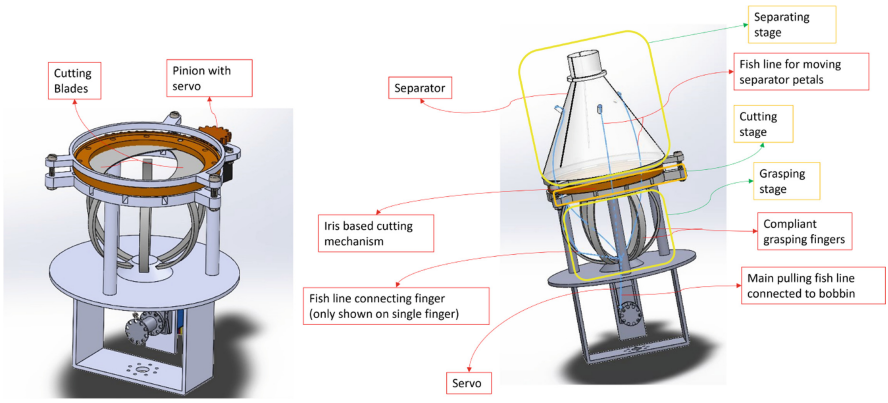


Fig. 1. 3D CAD model of the mechanical design of the complete gripper assembly built is shown without and with separator

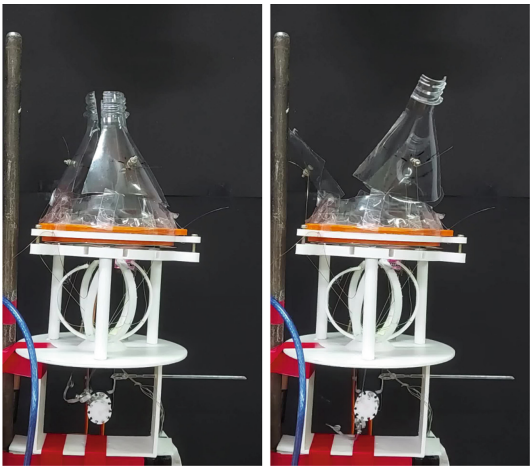


Fig. 2. Figure shows the working of separator mechanism

of damage. Instead, if a soft compliant, and adaptive structure of the gripper is used then there are almost negligible chances of damage. Currently, the existing state of the art of grippers uses various combinations of pulling and rotating motions for picking [21] which may damage the crop due to an inefficient way of operating. Hence, here a novel soft gripper (Fig. 1) attached with separator and iris pedicel cutting mechanism is proposed. The proposed gripper can separate the target tomato from the dense cluster and can cut the pedicel of the plant in addition to the gentle grasp in the final stage. It consists of six fingers in the form of curved compliant beams which forms a 360° open cage. The individual finger is attached with a fish line at the tip which passes through two connecting points on it at the midways of its curved surface. All the other ends of the wires

connecting the fingers are tightly connected between two nuts running over a common pulling bolt attached at the midway of the pulley in the form of a bobbin connected at Orange OT5325M servo motor directly. This on rotation pulls the fish line to close it to have a gentle grasp while elastic action of the beam will retract it to normal position when the motor rotates in the opposite direction to open the gripper cage. An Iris pedicel cutting mechanism with sharp curved mild steel blades is connected at the entry face of the gripper using three cylindrical rods which prune the tomato and can be picked up by the gripper. The cutting action of the blades can be controlled using a spur gear which is driven by a pinion attached to an MG90D micro servo motor. The cutting mechanism is installed with a conical frustum-shaped separator (Fig. 2) at its outer face with three petals that can open and close while taking the target tomato inside by separating it from the rest other from the dense cluster by applying a gentle force. Opening and closing actions can be controlled through the pulling action of a fishing line using a servo motor (shown blue in colour). This type of gripper configuration is found to be suitable for three different operations while picking tomatoes i.e. separating the target tomato from the dense cluster, cutting the pedicel, and finally grasping to hold the tomato in place without falling outside. Initially, the separator makes contact with the target tomato in a completely closed position, and then it starts moving its opening section gradually towards it to open only that much which allows the target tomato to enter into it leaving other tomatoes to slide over the conical surface to separate without causing damage to the target one. Also, it was tested experimentally that the gripper comes out of the cluster picking the target tomato gradually while the other tomatoes will not get damaged due to the collision between them as the movement will be very slow. This is in the initial approach however improvements will be made in the future. At this stage, we are not using any kind of force feedback for detecting the contact of the tomatoes with the conical surface of the separator to control the collision of the surrounding tomatoes however a gradual outward movement can be provided to the gripper to prevent any kind of damage.

The objective was to make a soft gripper with a cage-like structure using compliant fingers however an optimization scheme based on evolutionary computation such as a genetic algorithm is required to choose the optimal number of fingers for the firm stable grasping to hold the tomatoes firmly in place which is not part of the current work.

2.1 Kinematic Analysis

Since each finger is in the form of a curved compliant beam, the kinematic analysis of the gripper is done by the analysis proposed by [22] where the circular curved beam was analyzed and the tip deflection is studied using the modified Timoshenko beam theory (5) under the action of combined loads. Due to the symmetrical shape of the gripper, this analysis will apply to all other fingers. Using the motion capture through a high-resolution camera various kinematic parameters of the finger can be obtained as shown in Fig. 4. b which can be compared with the same parameters obtained through mathematical equations.

A relative comparison of the Fingertip position, velocities, angular position, and trajectory can be seen in Fig. 5. Through the path traced by the fingertip the workspace of a single finger can be obtained which can finally give the work volume. Bending moment and the shape which are the functions of the parameter **t** defined along the beam arc length are given by (1) and (4) while (6) and (7) gives x and y coordinates of the tip at any point on the beam respectively.

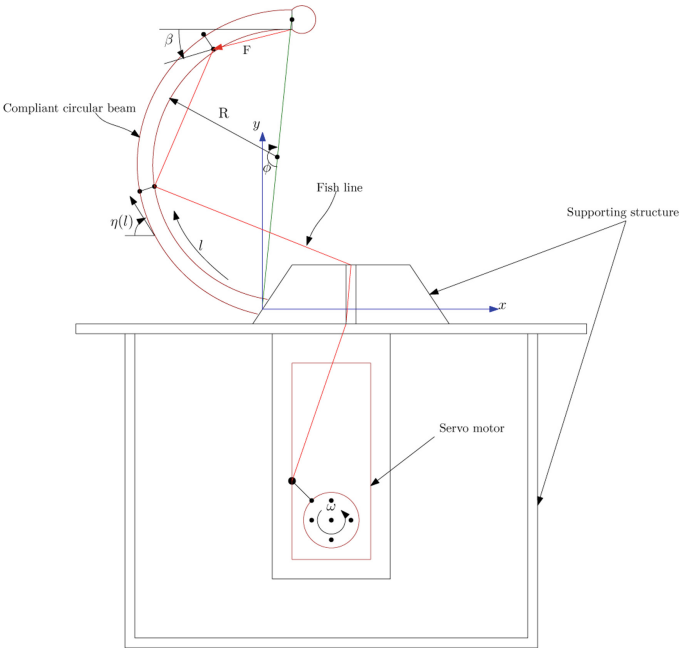


Fig. 3. Figure showing the actuation of a single finger acted upon by a force on its tip

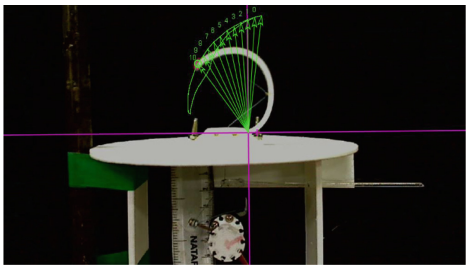


Fig. 4. Figure showing finger tip motion capture

$$N(l) = F \cdot [X \cdot \sin\beta - Y \cdot \cos\beta] \tag{1}$$

$$where, \quad X = x_d - x \quad and \quad Y = y_d - y \tag{2}$$

$$\eta'(l) = \frac{N(l)}{EI} + \frac{1}{R} \tag{3}$$

$$\eta''(l) = q \cdot (y' \cos\beta - x' \sin\beta) \tag{4}$$

$$x' = \cos\eta, \quad y' = \cos\eta \tag{5}$$

$$\eta''(l) = \frac{d^2\eta}{dl^2} = \zeta \sin(\eta - \beta), \quad l \in [0, L] \tag{6}$$

$$\zeta = qL^2, \quad \gamma = \frac{1}{R} = \frac{\phi}{L}, \quad q = \frac{F}{EI} \tag{7}$$

$$x_t = \int_0^L \cos\eta(l) ds \tag{8}$$

$$y_t = \int_0^L \sin\eta(l) \tag{9}$$

where the finger tip coordinates in deflected position are (x_d, y_d) , ϕ is the compliant finger arc angle at its centre considered to be 180°, β is tip force angle with the horizontal , η is the slope at a particular point and γ is the curvature as shown in Fig. 4.a. While, $\beta \in [-90°, -90°]$, $\zeta \in [-1, 1]$, $I = 15.75$ mm^4, $L = 100$ mm and Young's modulus, $E = 2410$ N/mm^2.

Using the camera the motion was captured to obtain the fingertip data from which we can get the following relationship of position with the motor angular rotation ω using curve fitting in MATLAB with R^2 value of 0.9853 and 0.9914 for both respectively which can be written as below:

$$\begin{aligned} x_t &= 204.3\sin(0.04136\omega - 0.29) + 132.6\sin(0.054\omega + 7.608) \\ &\quad +37.15\sin(0.07658\omega + 13.9) + 1.662\sin(1.617\omega + 2.081) \\ y_t &= 78.85\sin(0.01806\omega + 6.134) + 6.84\sin(0.1412\omega - 9.167) \\ &\quad +11.02\sin(0.1814\omega + 1.083) + 6.773\sin(0.1995\omega + 8.028) \end{aligned}$$

These expressions represent the variation in fingertip x and y positions as a function of servomotor angular rotation ω. As the bobbin connected to the motor rotates, we can see how it affects the single finger position, which is also true for the other fingers because the situation is symmetric due to its cage-like structure.

3 Experimental Evaluation

To evaluate the performance of the gripper assembly four sets of experiments were conducted. In the first lab trial, the testing of the tomato separator was performed to separate the target tomato from the dense cluster with complete gripper assembly as shown in Fig. 6. The average estimated time for separating

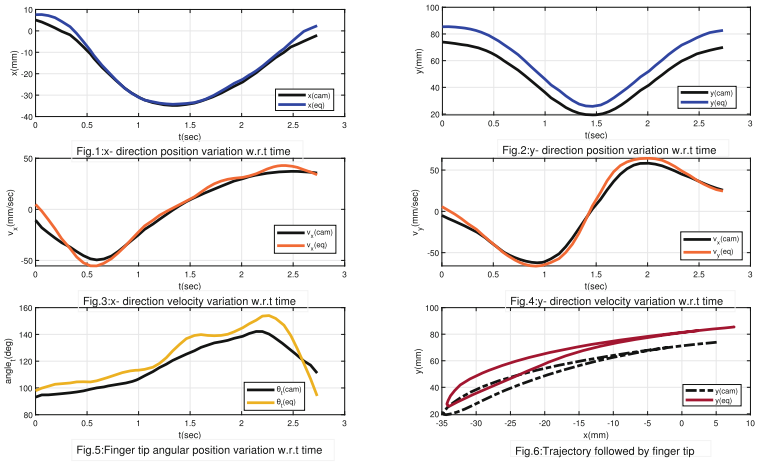


Fig. 5. Figure shows the variation of the finger-tip's x, y direction position, velocities, angular position with time and also the trajectory followed during a grasping cycle (open and close) obtained through mathematical expressions and motion capture respectively.

was found to be around 16.908 s for 10 different diameter samples of tomato. In the second experiment the test is conducted to evaluate the performance of the Iris pedicel cutting mechanism as shown in the Fig. 7.a. To test the cutting ability of the cutter the tomato with its pedicel was brought inside the cutting area up to a particular pedicel level then the sharp mild steel blades comes into action to make a fine cut. This experiment was able to complete the cutting operation in an average time of around 8.69 s for 10 samples of different diameters. The third experiment involves the testing performance of the gripping part of the assembly to evaluate the grasping capability of real tomatoes as shown in Fig. 7.b. An average estimated time of around 3.21 s for 10 samples of tomatoes of different diameters was found for grasping and holding. In the fourth experiment, the lab performance of the complete gripper assembly with separating and cutting mechanism along with the grasping part was tested. The tomatoes in the form of a dense cluster were brought near to the opening of the separator from the vertically downward direction pointing towards the target tomato as shown in Fig. 7.c. Due to the inclined surface of the separator, the tomatoes other than the target one slide over it constraining only the target tomato to move inside the gripping area. After moving inside up to a particular pedicel level the iris cutting mechanism comes into the action to make a smooth cut on it. Finally, the tomato will fall inside the grasping area consisting of the compliant soft fingers which will hold it firmly in its place. This complete operation took around an average of 28.81 s.

The experimental data was collected and for 10 samples variation of separating,cutting,grasping and total time with respect to diameter are presented in

Fig. 8. During the picking operation and also a comparison can be seen in the form of a bar graph shown in Fig. 9. This comparison is made between separating, cutting, grasping, and total time respectively for tomatoes of diameters ranging from 3.2 cm–6.4 cm, shown along the horizontal axis while the time is taken along the vertical axis. The observations show that as the diameter of the tomato sample increases, the cutting time increases due to the increment in the pedicel nominal diameter as the iris blades will take more time due to more resistance to cutting. However, the grasping time shows the opposite trend during grasping the sample tomato as the gripper fingers have to make less movement. The total time increases continuously from 26.66 s to 30.84 s as the sample diameter varies from 3.2 cm–6.4 cm.

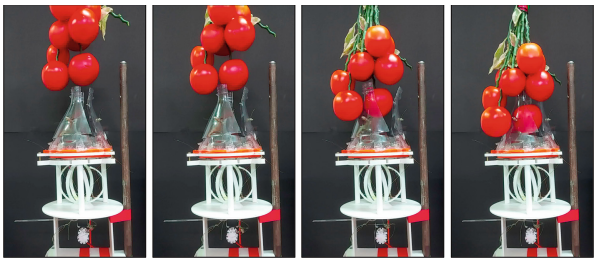


Fig. 6. Shows the experiment- 1: target tomato separating test

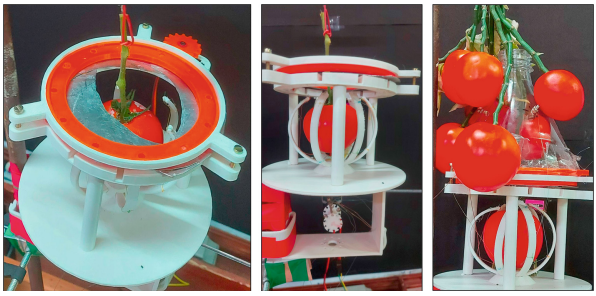


Fig. 7. Shows experiment-2, 3 and 4 i.e. pedicel cutting, grasping and complete operation test respectively.

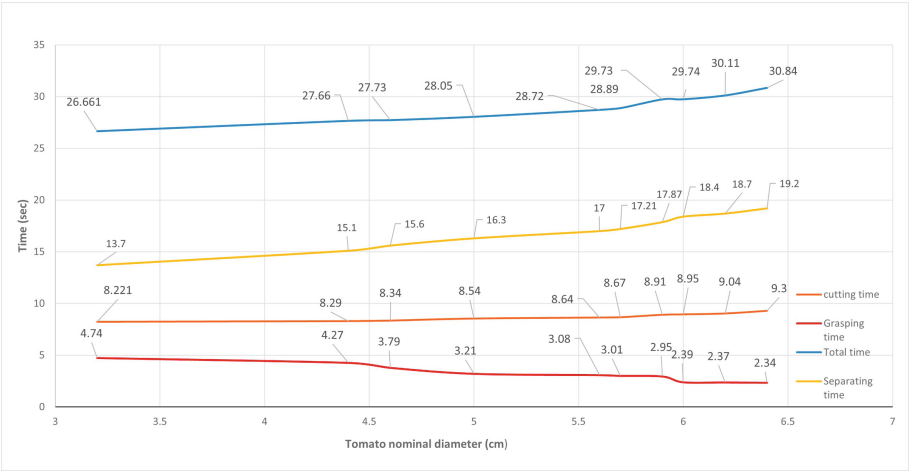


Fig. 8. Figure shows the plots for the variation of cutting, grasping, separating and total time respectively with the diameter of tomato sample while performing the picking operation

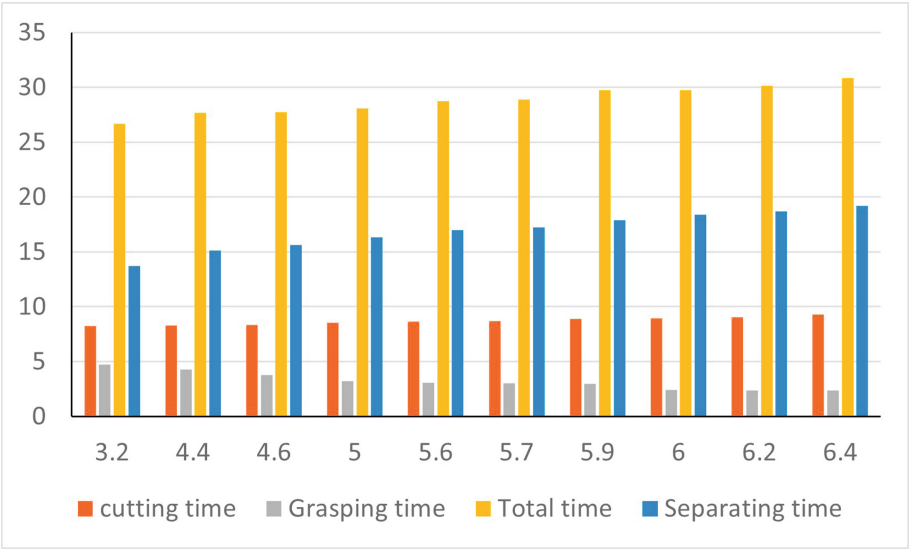


Fig. 9. Figure shows the bar graph for the variation of cutting, grasping, separating and total time respectively with the diameter of tomato sample while performing the picking operation

4 Conclusion

In this paper, a proof of concept system for a novel under-actuated soft robotic gripper with tomato separator and iris stalk cutting mechanism is presented. For its functioning, it employs thermoplastic polyurethane (TPU) based shape adaptive soft fingers in the form of curved compliant beams which can prevent any kind of damage and apply a required amount of force to grasp and hold the tomatoes. The main novelties of this system are the use of a tomato separating mechanism which separates the target tomato from the dense cluster and a mechanical iris pedicel cutting mechanism attached at the entry face of the gripper area driven by an MG90D micro servo actuator using a wheel and pinion mechanism. The functionality of this system has been estimated and evaluated quantitatively. The obtained results will form the basis for the development of the next level gripper with the capability of picking target tomato from the cluster using separating and cutting operations in the agriculture sector. Also, in future, we will develop a gripper with all its components i.e. separator, cutting, and grasping mechanism actuated by an actively cooled shape memory alloy actuator (Smart Material) to make it more compact and lightweight due to the high power to weight ratio. In addition, a new design of gripper fingers will be proposed using AI-based Generative Adversarial Networks (GAN) models.

References

1. Zhang, H., Wang, M.Y., Chen, F., Wang, Y., Kumar, A.S., Fuh, J.Y.: Design and development of a soft gripper with topology optimization. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 6239–6244. IEEE (2017)
2. Davidson, J.R., Silwal, A., Hohimer, C.J., Karkee, M., Mo, C., Zhang, Q.: Proof-of-concept of a robotic apple harvester. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 634–639. IEEE (2016)
3. Shintake, J., Cacucciolo, V., Floreano, D., Shea, H.: Soft robotic grippers. Adv. Mater. **30**(29), 1707035 (2018)
4. Tahir, A.M., Zoppi, M., Naselli, G.A.: PASCAV gripper: a pneumatically actuated soft cubical vacuum gripper. In: 2018 International Conference on Reconfigurable Mechanisms and Robots (ReMAR), pp. 1–6. IEEE (2018)
5. Yamaguchi, N., Hasegawa, S., Okada, K., Inaba, M.: A gripper for object search and grasp through proximity sensing. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1–9. IEEE (2018)
6. Sakamoto, N., Higashimori, M., Tsuji, T., Kaneko, M.: An optimum design of robotic hand for handling a visco-elastic object based on maxwell model. In: Proceedings 2007 IEEE International Conference on Robotics and Automation, pp. 1219–1225. IEEE (2007)
7. Yaguchi, H., Nagahama, K., Hasegawa, T., Inaba, M.: Development of an autonomous tomato harvesting robot with rotational plucking gripper. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 652–657. IEEE (2016)

8. Shah, S.H., Arsalan, M., Khan, S., Khan, M.T., Alam, M.S.: Design and compliance control of a robotic gripper for orange harvesting. In: 2019 22nd International Multitopic Conference (INMIC), pp. 1–5. IEEE (2019)
9. Lehnert, C., English, A., McCool, C., Tow, A.W., Perez, T.: Autonomous sweet pepper harvesting for protected cropping systems. IEEE Robot. Auto. Lett. **2**(2), 872–879 (2017)
10. Hayashi, S., Shigematsu, K., Yamamoto, S., Kobayashi, K., Kohno, Y., Kamata, J., Kurita, M.: Evaluation of a strawberry-harvesting robot in a field test. Biosys. Eng. **105**(2), 160–171 (2010)
11. Xiong, Y., From, P.J., Isler, V.: Design and evaluation of a novel cable-driven gripper with perception capabilities for strawberry picking robots. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 7384–7391. IEEE (2018)
12. Mghames, S., Hanheide, M., Ghalamzan, A.: Interactive movement primitives: planning to push occluding pieces for fruit picking. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2616–2623. IEEE (2020)
13. Van Henten, E., Van't Slot, D., Hol, C., Van Willigenburg, L.: Optimal manipulator design for a cucumber harvesting robot. Comput. Electron. Agric. **65**(2), 247–257 (2009)
14. Xiong, Y., Ge, Y., Grimstad, L., From, P.J.: An autonomous strawberry-harvesting robot: design, development, integration, and field evaluation. J. Field Robot. **37**(2), 202–224 (2020)
15. Monta, M., Konda, N., Ting, K.: End-effectors for tomato harvesting robot. Artif. Intell. Biol. Agri. **12**, 1–25 (1998). https://doi.org/10.1023/A:1006595416751
16. Bachche, S., Oka, K., Sakamoto, H.: Development of current based temperature arc thermal cutting system for green pepper harvesting robot. In: Proceedings of the Shikoku-section Joint Convention of the Institute of Electrical and related Engineers, Takamatsu, Japan, vol. 29 (2012)
17. Bachche, S., Oka, K.: Performance testing of thermal cutting systems for sweet pepper harvesting robot in greenhouse horticulture. J. Syst. Des. Dyn. **7**(1), 36–51 (2013)
18. Kondo, N., Yata, K., Iida, M., Shiigi, T., Monta, M., Kurita, M., Omori, H.: Development of an end-effector for a tomato cluster harvesting robot. Eng. Agric. Environ. Food **3**(1), 20–24 (2010)
19. Oktarina, Y., Dewi, T., Risma, P., Nawawi, M.: Tomato harvesting arm robot manipulator; a pilot project. In: Journal of Physics: Conference Series, vol. 1500, p. 012003. IOP Publishing (2020)
20. Ficken, L.A.: Iris mechanism. US Patent 4,804,108 (1989)
21. Navas, E., Fernández, R., Sepúlveda, D., Armada, M., Gonzalez-de Santos, P.: Soft grippers for automatic crop harvesting: a review. Sensors **21**(8), 2689 (2021)
22. Venkiteswaran, V.K., Su, H.J.: Pseudo-rigid-body models for circular beams under combined tip loads. Mech. Mach. Theory **106**, 80–93 (2016)

The Impact of Intelligent Objects on Quality 4.0

Sergio Salimbeni[1](✉) and Andrés Redchuk[2]

[1] Universidad del Salvador, C1051AAB Buenos Aires, Argentina
sergio.salimbeni@usal.edu.ar
[2] Universidad Rey Juan Carlos, Madrid, Spain

Abstract. The purpose of this work is to present the relationship that exists between Industry 4.0 (I4.0), smart objects and quality management, and how could benefit companies, users and the environment, based on interoperability and a horizontal and vertical integration of an entire value chain. The methodology was a mapping and coding method with a systematic literature review which was carried out to explore the current state of smart objects, quality management and I4.0. Findings show how enabling technologies of I4.0 turns quality management into a new data-driven one, named Quality 4.0. Based on a case study regarding a consumer electronic industry and with the use of a simple computational simulation, it is introduced the possible interaction between suppliers, manufacturers, and end users. One of the most significant traits of I4.0 is the possibility of interconnecting the successive phases of the complete product life cycle, from its conception to its disuse. For that purpose, all objects must be, at a minimum, identifiable for traceability. The security and confidentiality factor are vital for its operation. The originality of this work is the representation of an end-to-end integration using a computational simulation that uses discrete events, dynamic systems, and agent-based models in order to expose a total quality management digitalised system.

Keywords: Industry 4.0 · Quality 4.0 · Intelligent products · Intelligent objects · Digital Twin

1 Introduction

Competition between companies and the improvement of new technologies generate new quality management systems; because of that, enterprises seek to be agile, flexible and with dynamic capabilities in this Fourth Industrial Revolution (4IR), [1].

Quality is a distinctive and fundamental attribute for both, companies and consumers. In the first one, because it is important to optimise their operations and increase the demand for their products based on competitive and comparative advantages; in the latter, because consumers foresee their expectations to be met and, even, exceeded [2].

One of the principles established by William Deming, one of the great specialists in quality management, inferred that, in order to be competitive and stay in business, one should persevere in the continuous improvement of products and services. For his part, Joseph Juran, brought a more human dimension to quality management, dealing

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 287–298, 2023.
https://doi.org/10.1007/978-3-031-16281-7_28

with, among other factors, employee training and recognition [3]. In the past, concepts of quality were applied primarily to tangible products, but, due to the importance of service companies, quality also began to be applied to this vertical market segment and that is the reason why today quality is a key factor for any type of organisation [4].

Quality has evolved from Inspection through Control, Assurance, Total Quality Management (TQM) and, finally, Design for Quality, but, in recent years, the arrival of I4.0 enabling technologies has produced a significant leap in the evolution of quality [5].

The horizontal integration through the Value Stream, which is one of the pillars of I4.0, offers new opportunities for process optimisation, product functional compliance and, even, new sustainability and environmental goals [6].

The possibility of collecting data from supplies, semi-assemblies or finished products, throughout the entire value flow in real time and accessible to all stakeholders, allows to elicit data, traceability, root-cause analysis, consumer behaviours and information about disuse, discard or recycling of end products. This information is share between different departments of the company, Marketing, R&D, Production, Logistics, and that is known as vertical integration, being the conjunction between Information Technologies (IT) and Operational Technologies (OT). Such interoperability, between IT and OT, besides the horizontal integration are an important point of the I4.0 foundations [7].

The question that arises in this work is how the principles of I4.0, throughout the entire Valued Flow Life cycle, can have a positive impact on quality management in the Consumer Electronics Industry. Both, intelligent components, and intelligent end products, benefiting not only the company but its employees, end users and the hole environment, too. The connection between enabling technologies, vertical and horizontal integrations (VHI), intelligent products, and new human-centered concepts would be what gave form to this new generation of quality management the so-called Quality 4.0 (Q4.0). The quality management of the value stream in the 4.0 era entails disruptive innovations in the way of collecting data, analyse and predict behaviours. The use of this information and knowledge for decision making affects all areas of the company and the value chain. Knowledge of the status of the product in operation in the end user can be utilised for continuous improvements in processes, materials and the product itself, offering new after-sales services to the customer.

2 Research Methodology

A systematic literature review was first performed. The project work has been performed from January to December 2021. Open coding was used for data review and organization [8]. Mapping method and bibliometric techniques were used. A screening of 143 documents was carried out, 60 of which were selected based on the word “Quality4.0” for a deeper analysis. It was done as follows: (1) in academic databases (Scopus, IEEE Xplore, JSTOR, Researchgate and Scielo) with a search string through the combination of the operator “or” between keywords, the references that met the selected criteria were collected; (2) were published in conference proceedings, journal articles, magazines, book chapters and books between 2014 and 2022; (3) contained, at least, one of the search terms in the title, abstract, and/or keywords; (4) duplicates were removed; (5) those ones that did not have full texts available were discarded; (6) documents that defined I4.0,

Q4.0, Intelligent products, Intelligent objects and Digital Twin (DT) outside the scope of this research work were excluded; (7) they were classified according to the research question; (8) the collected documents were analysed and the data of interest for the research question was collected. During a second phase, a computational model of an entire Value Stream Life Cycle in and intelligent Consumer Electronic environment and its impact on Q4.0 was performed.

3 Literature Review

3.1 The 4th Industrial Revolution, Digital Transformation, and Industry I4.0

The technological revolution that society is currently going through, commonly known as 4IR, has emerged with machines and processes that are capable of self-manage, that is, they work without the need of human labour [9]. The major principles of the 4IR are applicable to the manufacturing industry, generally known as I4.0, which necessarily requires the total digitalisation of the value chain. The main objective of I4.0 is to develop a highly flexible model to produce personalised digital products and services allowing interaction between people, products and devices in real time [10]; that is possible thanks to the digitalisation of objects, components or any other type of agent, the so-called 'assets' in the I4.0 vocabulary (Fig. 1).

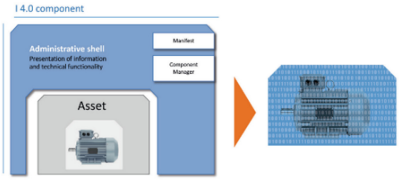


Fig. 1. Industry 4.0 component. Source: German Standardization Roadmap Industry 4.0, v.2020

While there is no doubt that Digital Transformation (DX) can bring many benefits, the process is not that simple and presents considerable challenges for companies, especially for small and medium-sized enterprises (SMEs) [11].

Digitalisation induces changes at different levels, such as: (1) at the process level: changes in the way work tasks and activities are carried out through the implementation of new digital tools or the digitalisation of activities that were previously manual; (2) at the organisation level: new ways of providing existing or new offerings; (3) at the business level: changing roles and value chains and (4) at the societal level: redesigning types of work, competition, and infrastructure for digitalisation [12].

Because digitalisation deals with information processing and everything can be turned into digital information, it affects all human activities. That is why DX implies a series of changes. Based in an study carried out by [13], stakeholders highlighted the following three challenges: (1) the need to develop and update digital skills within companies, public administration, organisations and education; (2) the need to continuously adapt to new digital technologies with state-of-the-art infrastructures and services and

(3) the need to develop medium to long-term synergies and innovative technological solutions involving different actors from the public and private industry.

There are technologies that have been evolving in recent years and others that are truly innovative. The novelty is, nowadays, the combination and association of those technologies, which provide the means to generate great leaps in terms of performance and user capabilities. They are called enabling technologies and are the facilitators of I4.0. [14] established that there are 9 enabling technologies: (1) Industrial Internet of Things (IIoT), (2) Big Data and Analytics (BDA), (3) Horizontal and Vertical System Integration (HVSI), (4) Cloud Computing (CC), (5) Additive Manufacturing (AM), (6) Virtual and Augmented Reality (VAR), (7) Autonomous Robot (AR), (8) Computational Simulation (CS) and (9) Cybersecurity. Other authors also consider Artificial Intelligence (AI) as another important enabling technology [15].

There is no doubt that I4.0 brings a series of benefits. In an study performed by [16], the author states that I4.0 increased profitability and product quality, diminish maintenance, improve competitive advantage, reduce industrial waste, provides economic welfare of the society and increase return on investment and customer satisfaction. For this to succeed must be fulfilled the VHI from end to end throughout the entire value chain (Fig. 2) [17].

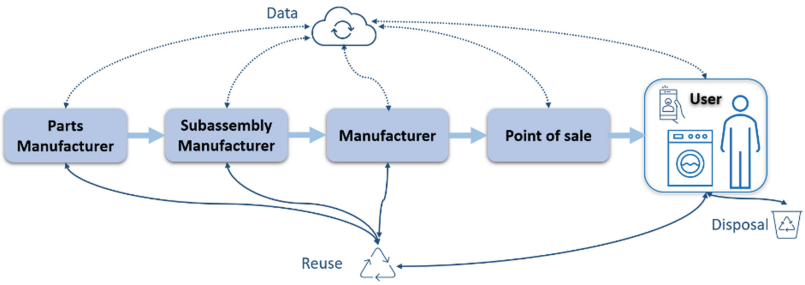


Fig. 2. Value flow. Source: author's own

It may be concluded that there is a positive impact on quality management due to I4.0 added to the data processing capacity and the collaboration of AI, especially given by the possibility of monitoring, control, and elicitation of large volumes of data at the production, commercial and end users' stages; but, for all this to be possible, intelligent components are needed.

3.2 Smart Objects

A detailed study carried out by [18] proposes an smart objects classification according to the level of "intelligence". As defined by the author, they can be classified into: (a) Digital, (b) Connected, (c) Responsive and (d) Intelligent.

Basically, Digital Objects are ones equipped with IT. Connected Objects have got a unique identification, identity, uniquely identifiable and they are human readable. Responsive Objects are characterized by the detection, data collection, knowledge of

the context in real time and situationally. They can adapt and react to changes in its environment in the form of a stimulus-response way. And finally, the Intelligent Objects, those that can make decisions about themselves and their interactions with external entities; they have autonomy and self-management and act intelligently and independently; it is said to have proactivity, it means, action against a future situation. An intelligent product is an entity that provides greater simplicity through proactive, actuate beforehand, behaviour.

3.3 Quality 4.0

The term 'quality' is used in different ways and has different meanings for different people, professions, and disciplines. It is, sometimes, a very subjective concept because of the various approaches and philosophies. For instance, TQM, zero-defect concept or quality management principles defined by DIN in ISO 9000.

The authors [19] define Quality as a functional management discipline that seeks for business processes and workflows optimization, considering physical and time constraints, as well as preserving the features of products or services and their subsequent use and growth. The author [20] summarises them up in four concepts: "quality planning, quality control, quality assurance and quality improvement" and their relation to effectiveness, efficiency and continuous improvement. On the other hand, it is proper to keep in mind that TQM can be viewed from two perspectives: "soft and hard practices". The soft practices of TQM are theoretical aspects related to human resources and management principles and concepts, while the hard practices refer to tools and techniques.

The first mention of the term 'Quality 4.0' was in 2015 in the Report on the future of quality of the ASQ [21], which established that an evolution was expected in the techniques, tools and methods of quality in factors such as: quality as inspection, quality as design, quality as empowerment and quality as discovery, an adaptable and intelligent environment to solve challenges and problems. For those reasons, quality professionals should now use new knowledge such as systems thinking, data-driven decision making, leadership for organisational learning, and understanding how decisions affect people, lives, relationships, communities, wellbeing, health and society in general.

A comprehensive treatment by means of collecting data from many different sources with BDA will become the standard to support real-time decision making [22]. Information security is a must in these cases because of the massive use of data. Although it is beginning to be used scalable blockchain-empowered systems several challenges have to be faced, such as consensus algorithms and privacy protection needs in manufacturing systems [23].

While I4.0 evolves from automation to smart manufacturing, Q4.0 moves from a process approach to a customer-centric data-driven approach. In accordance with that, [24] emphasise that Q4.0 could be defined as the adaptability of a product at any stage of its life cycle to meet customer needs, taking into account the interest of other stakeholders along the value chain.

From a wider point of view, Q4.0 could be taken as a new method by which digital tools can be used to drive improvements across the value chain [25], and it is also said that quality shifts from production to system design [26].

In regard to change management, the key factors for the implementation of Q4.0, according to [27], are: handling big data, prescriptive analysis, VHI, leadership, training, organisational culture and senior management commitment.

In addition, a reference framework belonging to [5] establishes the key axes of the future of Q4.0 and it will be discussed latter in the case study. They are: (1) Quality as a data driven discipline, (2) the application of modelling and simulation for evidence-based quality engineering, (3) health monitoring and prognostics for quality, (4) integrated quality management, (5) maturity levels with respect to the 4IR, (6) integrating innovation with quality and managing for innovation, (7) Quality 4.0 and data science, (8) integrating reliability engineering with quality engineering, and (9) information quality.

Further, taking into account some other factors such as leadership, competence and human-centered strategies would lead to what is already being called Quality 5.0 [28].

Another essential topic to keep in mind is the importance of data´s quality. VHI and systems interoperability multiplies the information channels and the amount of data that is collected (Fig. 1) and that is one of the gist why BDA become so important for decision making in this new generation of quality management. Then, a data-driven approach to quality management enables successful improvement models to be implemented [1]. There are, however, some challenges and paradigms that must be overcome. First, frequently, manufacturers focus on the quality of their products, but unfortunately not onto Data Quality or the Data Quality Management processes, and second, the IIoT capability of manufacturing companies is not as high as expected [29]. Because of that, it is very important to pay great attention to the treatment and care of data quality. There are mostly two ways to improve the quality of data: (a) data-centric and (b) process-centric approaches. The data-centric approach is preferable when it is necessary to correct randomly generated data non-conformities, whilst process-centric approach can be employed to the systematic identification of root- causes of the data nonconformities. It also allows the execution of specific procedures to remove such root-causes through continuous improvement of the processes that include people, procedures, methods, tools and equipment. In process-centric approaches, it is necessary to assess how well the data quality management processes have been performed. There are also different types of data and the treatment that each one requires may be different. The possible kinds of data used in such environments could be sensor data produced by sensors, observed metadata showing the characteristics of the sensor data, device metadata indicating the characteristics of the sensor device, business data for maintenance and service, and external data such as weather and traffic information.

Standards which deal with data quality could be found, although they were not initially intended for IIoT systems. For instance, ISO 8000 defines data quality principles, fundamental concepts and data quality vocabulary whilst ISO 8000-6x series addresses data quality management processes [30].

Coming back to data collection and analysis, it can be said that while many traditional data analysis techniques can be used to develop quality processes and products, the massive data sets collected by I4.0 technologies need to be carefully mined with analysis methods based on BDA technologies. Effective decisions can be made from such information, including, for instance, Lean Six Sigma cycles. The use of data analysis methods at each stage is of critical importance in making decisions [31]. On the other hand, [32]

state that I4.0 is based on the use of data to conform intelligent processes and systems, enhancing the way in which companies make decisions at all levels. However, it should be noted that the myriad natural variations in processes require significant additional work to realize the full potential of smart manufacturing.

The DX of companies, based on enabling technologies, offers several new opportunities. Data capture and handling of large data volumes allow different scenarios to be analysed, and processes and products could be improved, but great care must be taken to preserve the quality of said data in order not to generate new and bigger errors.

4 Case Study

This case study analyses the impact on quality management throughout the life cycle of the value stream in an intelligent consumer electronics factory, as defined in RAMI4.0, starting from the customer requirements, going through the design, production, distribution, use and recycling of the product. This new approach in quality management is the so called Q4.0. The aim of this work is to show how, based on the I4.0 concept, the data elicitation, its analysis, information and knowledge could be very useful for all the stakeholders and for several departments of the company providing information to the end user, too. The fact that communications was bidirectional, gives the opportunity to perform both, monitoring and control of digitalised intelligent assets. Although there is several commercial software specialized in the supply chain, Anylogic™ was chosen in this case for two reasons: the flexibility in the use of its different simulation methods, and because of its free Personal Learning Edition version.

Since there is not a complete implementation like this example in the real world, this study is made through a CS model paving the way for a future DT. It must be kept into account that DTs are made up of three components: (1) physical world, (2) the virtual models, and (3) the link between both [33]; therefore, a second phase of this research work would be missing for the link between both, real and virtual worlds. Before making a description of the model it must be remembered that there are three kinds of CS: Discrete Events (DE), System Dynamics (SD) and agent-based (AB) models. The DE modelling takes a process-oriented approach; SD are represented as a sequence of operations performed on entities and in case of AB models the system is described from the point of view of individual objects which could interact with each other and with the environment [34]. The three methods are used for this case study. Basically, the DE is used for the core process of the value system, and a SD and AB model to assess consumer behaviour (Fig. 3). It should be noted that in this study only the logical modelling has been carried out, and neither a 2D nor a 3D simulation was deployed yet.

This simplified model portrays a washing machine manufacturing process with three different suppliers of raw materials or sub-assemblies to the finished product ready to be delivered.

The end-products go to the points of sales, which keep them in inventory to be sold. Sales depend on the marketing activities, the availability of the product in a timely manner, and the rate of adoption by potential buyers, who are influenced by advertising and "word of mouth". The quality of the manufacturing process translates into maintaining competitive costs and availability of products to be delivered to the points of sale

Fig. 3. Computational simulation model with Anylogic™ software. Source: author's own

in a timely manner. Sales and production plans come from estimating and forecasting demand. The quality of the product and after-sales service benefits word of mouth and customer loyalty. The information on behaviour and regularity of use by the consumer, allows the development of new post-sale services, such as, for example, prescriptive maintenance. On the other hand, the traceability of the product until its disuse, not only benefits the care of the environment, but also allows the implementation of a circular economy.

As stated, modelling that attempts to approximate real-world processes to important punctual events is called DE modelling and it is mainly used, as in this case of washing machine manufacturing, for those process-focused cases and are represented through process flow diagrams.

SD is primarily used for long-term analysis. It can represent people, products, events, and other discrete items. It is very important in this kind of systems the flows and delays [35], and because of that, it was selected to model consumer buying behaviour.

Finally, AB modelling has been combined with SD to model people and potential washing machines buyers in this case study, because it is an essentially decentralized and as opposed to system-level approach. Active agents are identified, and their behaviour is defined.

One of the novel advantages of this model would be, for example, that the Marketing Department could know how often the washing machine is used, know consumer preferences based on the day of the week and even the weather. The end user may also be advised, as an after-sales service, with the best day and time for washer machine use based on the hours of lowest electricity consumption with its corresponding benefit in lower rates. Also as tips for use and maintenance depending on the intensity and mode of use. On the other hand, the R&D department can know the operating characteristics and wear of parts based on the monitored variables. In our model, among other magnitudes, temperature, torque, vibration, temperature and humidity are monitored as seen in Fig. 3.

5 Discussion

This simple model has represented the digitization and the horizontal and vertical integration of the entire value chain of the manufacturing and sales of appliances. As illustrated in Fig. 2, data about the preferences of use of said product can be collected and with it, areas of marketing, R&D could improve and adapt the design to the client's needs. Continuous monitoring, through internet, would allow BDA functions generating prescriptive maintenance information, offering the customer new after-sales services.

The focus of quality management is on components, processes and products. To that end, objects must have a certain level of intelligence, which allows, at least, geolocation and bidirectional communication.

Based on that simulation, sensitivity analyses could be carried out in order to optimize the processes achieving operational excellence. That was accomplished with time studies and activity-based cost (ABC) analysis, as well as nonconformity cost estimations.

Continuous monitoring and predictive and prescriptive maintenance improve the quality of the product and customer satisfaction.

Quality has been presented in this article as a data-driven strategy. The modelling and simulation application for evidence-based quality engineering, health monitoring and forecasting for quality, integrated quality management, the innovation of Quality Management 4.0 and data science, and the quality of information was introduced.

Great attention must be paid to the security and confidentiality of the user client's data, as well as to consider the ethical problems that could arise from its use.

6 Conclusion and Future Research

An introduction to I4.0, smart components and Q4.0 has been introduced. Based on a simple CS model, it was analysed how an integrated system of suppliers-manufacturers-users could share information to improve the quality of components, assemblies and products. For example, and among many other opportunities for improvement, based on prescriptive maintenance services, performance and customer satisfaction would be improved, who, indirectly and through word of mouth, would help the company increase its sales. For its part, the company would collect a huge quantity of data to analyse and in that way improve said product. Data-driven quality is possible and delivers a myriad of benefits for all stakeholders. The digitization of the entire value chain, horizontal and vertical integration, requires technologies that take care of information security and preserve the confidentiality of data, fundamentally those generated by end users. End-to-end traceability, another benefit of this integrated system, also offers the possibility of entering into a circular economy.

For future research work, a greater number of variables, including geolocation and scenario analysis will be required. A possible next step would be the integration of said CS into the physical world in order to develop a DT.

Taking into account other factors in quality management, such as leadership, competence, skills and human-centered strategies, it would lead to what is already being called Quality 5.0.

References

1. Sariyer, G., Mangla, S.K., Kazancoglu, Y., Ocal Tasar, C., Luthra, S.: Data analytics for quality management in Industry 4.0 from a MSME perspective. Ann. Oper. Res. no. 0123456789 (2021). https://doi.org/10.1007/s10479-021-04215-9
2. Carvalho, A.V., Enrique, D.V., Chouchene, A., Charrua-Santos, F.: Quality 4.0: An overview. Procedia Comput. Sci. **181**(2019), 341–346 (2021). https://doi.org/10.1016/j.procs.2021.01.176
3. Oakland, J.S.: Gerenciamento da Qualidade Total, 5ta ed. Sao Paulo: Nobel (1994)
4. Koontz, H., Weihrich, H., Cannice, M.: Administración. Una prespectiva global y empresarial, 13ra ed. McGraw-Hill/Interamericana editores s.a. (2008)
5. Zonnenshain, A., Kenett, R.S.: Quality 4.0—the challenging future of quality engineering. Qual. Eng. **32**(4), 614–626 (2020). https://doi.org/10.1080/08982112.2019.1706744
6. Gaiardelli, P., et al.: Product-service systems evolution in the era of Industry 4.0. Serv. Bus. **15**(1), 177–207 (2021). https://doi.org/10.1007/s11628-021-00438-9
7. Standarization-Council-Industry, DIN and DKE Roadmap. German Standardization Roadmap Industrie 4.0 version 4, 4th ed. Berlin: DIN (2020)
8. del, M., Hernández Sampieri, P., Fernández Collado, R., Baptista Lucio, C.: Metodología de la Investigación, 5ta ed. México, D.F.: McGraw Hill (2010)
9. Öztürk, F., Kayar, A.: Product lifecycle management in smart factories : Industry 4 . 0 Applications, In: 10th International Symposium on Intelligent Manufacturing and Service Systems, pp. 1420–1427 (2019)
10. Oluwaseun, A., Numbu, L.P.: Industry 4.0: The Fourth Industrial Revolution And How It Relates To The Application Of Internet Of Things (IoT). J. Multidiscip. Eng. Sci. Stud. **5**(2) (2019)
11. Rauch, E., Stecher, T., Unterhofer, M., Dallasega, P., Matt, D.T.: Suitability of Industry 4.0 concepts for small and medium sized enterprises: Comparison between an expert survey and a user survey. In: Proceedings of the International Conference Industrial Engineering Operation. Management, pp. 1174–1185 (2019)
12. Elg, M., Birch-Jensen, A., Gremyr, I., Martin, J., Melin, U.: Digitalisation and quality management: problems and prospects. Prod. Plan. Control **32**(12), 990–1003 (2021). https://doi.org/10.1080/09537287.2020.1780509
13. Brunetti, F., Matt, D.T., Bonfanti, A., De Longhi, A., Pedrini, G., Orzes, G.: Digital transformation challenges: strategies emerging from a multi-stakeholder approach. TQM J. **32**(4), 697–724 (2020). https://doi.org/10.1108/TQM-12-2019-0309
14. Mabkhot, M.M., et al.: Mapping industry 4.0 enabling technologies into united nations sustainability development goals. Sustainability **13**(5), 1–35 (2021). https://doi.org/10.3390/su13052560
15. Chouchene, A., Carvalho, A., Lima, T.M., Charrua-Santos, F., Osório, G.J., Barhoumi, W.: Artificial intelligence for product quality inspection toward smart industries: quality control of vehicle non-conformities, In: ICITM 2020 - 2020 9th International Conference on Industrial Technology and Management, pp. 127–131 (2020). https://doi.org/10.1109/ICITM48982.2020.9080396
16. Ammar, M., Haleem, A., Javaid, M., Walia, R., Bahl, S.: Improving material quality management and manufacturing organizations system through Industry 4.0 technologies. Mater. Today Proc. **45,** 5089–5096 (2021). https://doi.org/10.1016/j.matpr.2021.01.585

17. Batchkova, I.A., Gocheva, D.G., Georgiev, D.: Iec-62264 Based Quality Operations Management According the Principles of Industrial Internet of Things, In: Scientific Proceedings Xiv International Congress Machines. Technologies. Materials, vol. VI, no. 3, pp. 431–434, (2017). https://mtmcongress.com/proceedngs/2017/Summer/6/09.IEC-62264
18. Raff, S., Wentzel, D., Obwegeser, N.: Smart Products: conceptual review, synthesis, and research directions. J. Prod. Innov. Manag. **37**(5), 379–404 (2020). https://doi.org/10.1111/jpim.12544
19. Foidl, H., Felderer, M.: Research challenges of industry 4.0 for quality management. In: Felderer, M., Piazolo, F., Ortner, W., Brehm, L., Hof, H.-J. (eds.) Innovations in Enterprise Information Systems Management and Engineering. Lecture Notes in Business Information Processing, vol. 245, pp. 121–137. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-32799-0_10
20. Ali, K., Johl, S.K.: Impact of total quality management on SMEs sustainable performance in the context of industry 4.0. In: Al-Emran, M., Al-Sharafi, M.A., Al-Kabi, M.N., Shaalan, K. (eds.) Proceedings of International Conference on Emerging Technologies and Intelligent Systems. Lecture Notes in Networks and Systems, vol. 299, pp. 608–620. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-82616-1_50
21. Arsovski, S.: Social Oriented Quality : From no. 1, pp. 397–404 (2019)
22. Ates, E., Bostanci, E., Guzel, M.: Security evaluation of industry 4.0: understanding industry 4.0 on the basis of crime, big data, internet of things and cyber physical systems. Güvenlik Bilim. Derg. pp. 29–50, (2020). https://doi.org/10.28956/gbd.695889
23. Leng, J., et al.: Blockchain-empowered sustainable manufacturing and product lifecycle management in industry 4.0: A survey, Renew. Sustain. Energy Rev. **132** (2020). https://doi.org/10.1016/j.rser.2020.110112
24. Salimova, T., Vatolkina, N., Makolov, V., Anikina, N.: The perspective of quality management system development in the Era of industry 4.0. Humanit. Soc. Sci. Rev. **8**(4), 483–495 (2020). https://doi.org/10.18510/hssr.2020.8447
25. Sony, M., Antony, J., Douglas, J.A.: Essential ingredients for the implementation of Quality 4.0: a narrative review of literature and future directions for research. TQM J. **32**(4), 779–793 (2020). https://doi.org/10.1108/TQM-12-2019-0275
26. Dias, A.M., Carvalho, A.M., Sampaio, P.: Quality 4.0: literature review analysis, definition and impacts of the digital transformation process on quality. Int. J. Qual. Reliab. Manag. (2021). https://doi.org/10.1108/IJQRM-07-2021-0247
27. Dovleac, R.: Knowledge management systems in quality 4.0. MATEC Web Conf. **342**, 09003 (2021). https://doi.org/10.1051/matecconf/202134209003
28. Milunovic Koprivica, S., Maric, A., Ristic, O., Arsovski, S.: Social oriented quality: from quality 4.0 towards quality 5.0. Proc. Eng. Sci. **1**(2), pp. 405–410 (2019). https://doi.org/10.24874/pes01.02.038
29. Kim, K., Baker, M.A.: Luxury Branding in the Hospitality Industry: the impact of employee's luxury appearance and Elitism attitude. Cornell Hosp. Q. (2021). https://doi.org/10.1177/19389655211022660
30. Kim, S., Pérez-Castillo, R., Caballero, I., Lee, D.: Organizational process maturity model for IoT data quality management. J. Ind. Inf. Integr. **26**, (2022). https://doi.org/10.1016/j.jii.2021.100256
31. Dogan, O., Gurcan, O.F.: Data perspective of lean six sigma in industry 4.0 era: A guide to improve quality. In: 2nd International Conference on Industrial Engineering and Operations Management, Paris, pp. 943–953 (2018)

32. Oliff, H., Liu, Y.: Towards industry 4.0 utilizing data-mining techniques: a case study on quality improvement. Procedia CIRP **63**, 167–172 (2017). https://doi.org/10.1016/j.procir.2017.03.311
33. Qi, Q., Tao, F.: Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. IEEE Access **6**, 3585–3593 (2018). https://doi.org/10.1109/ACCESS.2018.2793265
34. Borshchev, A.: Multi-method modelling: AnyLogic, vol. 9781118349 (2014)
35. Ghosh, A.: Dynamic Systems for Everyone Understanding How Our World Works. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-10735-6

Mutual Resiliency and Lean Analysis for Supply Chain Management in a COVID-19 Ventilator Production Case Study

Alan Caraveo, Saikiran Krishnamoorthy, Jonah Marie Lawas, Parthasarathy Periaswamy, Majid Sodachi, and Omid Fatahi Valilai(✉)

School of Business, Social and Decision Sciences, Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany

{A.Caraveo,S.Krishnamoorthy,J.Lawas,P.Periaswamy,M.Sodachi, O.FatahiValilai}@Jacobs-University.de

Abstract. Lean management and resilience are studied as significant paradigms in the supply chain in unaggregated domain. As studied in the literature, the resiliency can contribute to the flexibility of supply chains for fulfilling the fluctuations by enabling redundancy. However, the level of redundancy can be an interesting topic from lean management perspective to be managed. This paper has focused on the context of aggregation of Lean and resiliency management these and has proposed a trade-off matrix method. The integration and aggregation of lean and resilient supply chain (SC) practices are investigated with a special focus for creation of synergy. The paper also has investigated the impact of this integration on SC performance through a case study in context of the COVID-19 ventilator production. This case has been designed to highlight the tradeoff analysis among lean and resilience concepts. By applying an Analytic Hierarchy Process model, the paper has enabled a decision-making framework for determination of redundancy level. The data collection for including and documentation of information about various sections of the supply chain is conducted to determine the optimal solution by service level evaluation.

Keywords: Supply chain management · Lean production · Supply chain resiliency · COVID-19

1 Introduction

Many companies today face various challenges brought about by disruptions. A few of the noteworthy disruptions that have occurred recently include Brexit, imposing taxes to China by the United States, and the current coronavirus pandemic [1]. The coronavirus is significantly affecting the supply chains of companies. While they try to problem-solve issues brought about by the pandemic, organizations also continue to face competitive operational pressures in terms of cost [2]. This means that lean, a strategical concept that reduces costs and optimizes profits [3], should be enforced by companies in their operations. Lean also focuses on detecting issues and remedying them [4], a reactive

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 299–309, 2023.
https://doi.org/10.1007/978-3-031-16281-7_29

approach that helps companies prevent further escalation. However, the rigidity that comes with implementing lean practices has been argued to leave companies in a state of inflexibility and less resilience [5] to sudden and impactful disruptions. Likewise, this argument is contradicted by Toyota's ability to quickly and with coordination respond to disruptions in supplies within its lean supply network [5]. These perspectives present a question as to whether lean and resilience are always at a trade off with each other [5]. There are multiple features within which both lean and resilience can be considered. Suppose a certain synergy or trade off can be realized from both the concepts. In that case, this could lead to valuable reference for establishing decisions that help mitigate expensive disruptions and create more seamless, efficient, and sustainable supply chains [3].

Taken in the context of small and medium-sized companies (SMEs), it can be noted that not many in-depth studies have been established for the implementation of lean in SMEs [6]. Alkhoraif et al. also highlight this in their study [7]. The same can be said for resiliency in SMEs wherein research with an SME focus is limited [8]. Looking into how lean and resilience play a role in keeping SMEs vigilant amidst globalization and disruptions could significantly contribute to SMEs' continuous business developments.

It is important to achieve the optimal combination of lean and resiliency as mentioned above and monitor and sustain it in the long-term by providing adequate consideration to barriers [9]. Barriers are forces that threaten the long-term targets of the firm [10, 11]. Uncertainty is a complicated phenomenon capable of disrupting any point across the entire supply chain network [12]. Several sources of uncertainty like demand spikes, technology changes, political instability, economic crisis, currency arbitrage, microeconomic stagnations, etc., can serve as potential agents of barriers [10]. There are several other potential inter-and intra-organizational barriers to resilient – lean implementation like lack of collaboration, increasing network complexities, lead time variabilities, supply disruptions, technology inaccessibility, and lack of information and financial capabilities [11]. It is mandatory to realize that the barriers to supply chain resiliency and lean are often common or interrelated to each other, reinstating the importance of arriving at an amicable trade-off between them for optimal business performance.

On the other hand, the enablers which negate the barriers are not separate but are in many ways interrelated to obstacles and the trade-off between lean and resiliency strategies In the past, several research papers emphasize the critical enablers are providing a competitive edge in business like sourcing flexibility, process flexibility, product flexibility, and transportation flexibility [13]. For instance, transportation flexibility provided a competitive edge to Chrysler over Ford to resume its delivery operations after the 9/11 terror attacks [11]. Companies frequently employ a risk management analysis to visualize the barriers and their impacts, consequently developing action plans to mitigate them [14].

This paper investigates the interaction and trade-offs of lean and resilience to reduce the negative effects of disruptions and help the SME firms handle the operational costs in this regard. It further discusses the issues that impede companies from implementing the said concepts and outlines how companies can leverage lean to build resiliency through an established trade-off matrix.

The paper is organized into different sections. In the next section, the paper discusses the dominant terminologies to establish a standard and concrete perspective. Then, the research gap is investigated. Considering all the elements of lean and resilience, a matrix based guidelines for trade-off analysis are discussed. From there, the barriers identified from the SME perspective is presented and related to the matrix. The trade-off matrix is then used in a case study to explain further how it can be utilized. Finally, the implications and conclusions are derived.

2 Literature Review

2.1 Lean Procurement

Lean is a strategical tool that focuses on the concept of eliminating wastes to create value through cost reduction, flexibility, and continuous process improvements [1]. The Lean Enterprise Institute further supports this definition while adding that lean creates value through the utilization of minimal resources [10]. The use of lean techniques encompasses the entire supply chain network, including information flow [3], and is not just about decreasing inventory level. Instead, it is about having just the right amount of inventory when and where needed, so processes run seamlessly [1]. Just-in-Time (JIT) and total quality management (TQM) are two of the more commonly used tools under lean.

On the other hand, procurement has been described by multiple authors [13] as the advancement of traditional purchasing where reduction of cost and low prices used to be the focus [14]. This means that procurement is now responsible for handling the complexities of purchasing decisions and managing suppliers and resources both through the internal and external perspectives of an organization. Current procurement processes are facilitated mostly by purchasing systems that help to identify the best supplier measured through criteria such as reliability, quality, and delivery performance [15].

Lean procurement, then, is the incorporation of lean into procurement practices. It is about getting 'the right things at the right time and in the right quantity' while minimizing waste and being flexible and adaptable [16]. Wilson and Roy (2009) further mention that lean procurement generally is characterized by small lot sizes and frequent purchases from a small pool of suppliers who also promote lean to their suppliers [15]. These descriptions can be denoted that lean procurement has a more external emphasis [4]. Lean companies tend to collaborate with suppliers for further improvement and development initiatives [10] and foster long-term relationships while buyers also help suppliers deal with challenges and concerns.

There are many views within which lean procurement can be considered. Based on the literature and the research direction, lean procurement can be described as an integrated and collaborative effort extending to the supplier's suppliers. It can incorporate different lean concepts (e.g., Kanban) but are generally characterized by reduced costs, reduced waste, small lot sizes, frequent deliveries, and few suppliers.

2.2 Resilience

Many research Studies from different perspectives have defined supply chain resilience. Some authors have defined resilience only from a proactive stance. For example, Carvalho et al. described resilience as the potential of supply networks to withstand unforeseen obstructions [3]. On the other hand, some authors consider resilience as a combination of both proactive and reactive strategies. Wieland and Wallenburg state that resilience is the optimal combination of agility, focusing on its adaptability and pace, and stability, focusing on proactivity and preparedness to subjugate disturbances [11]. Supply chain resilience is the supply network's transformative ability to organize, respond, and rebound from unforeseen events [14]. Some authors provided a comparative perspective to resilience, emphasizing the agility of the supply chain. For example, Juttner & Maklan defined resilience as the relative capability of specific supply chains to more adaptively rebound in the face of unavoidable threats than the others [11]. Several authors emphasized that resilient abilities should restore the system to its initial condition. Resilience is the potential of a company to absorb shocks, thereby positively affecting its performance by empowering the supply network to swiftly return to its normal operation state. Ates & Bititci defined resilience as the firm's ability to withstand and adjust itself concerning a violent disruption to maintain its business performance [11]. Some authors provided more optimistic and inclusive definitions that resilient abilities can restore the system to a state superior to its initial operating state. For example, Christopher & Peck defined resilience as a system's capability to rebound to its earlier state or a new beneficial state after a disruption [11, 13]. Supply chain resilience is the potential to adapt to unforeseen incidences swiftly, to be able to rebound to the original state or a more favorable state [13].

2.3 Literature Gap Analysis

The definitions and perspectives of lean procurement have been made clear through different research papers. Lean procurement, a value-adding concept based on the elimination of wastes [16], is focused on having quality, small lot sizes [4, 15] regularly ordered from a few suppliers who deliver at the right location in a just-in-time manner [10, 15, 16]. Lean procurement takes a more external standpoint wherein multiple commitments, efforts, and support [15, 16] to practice its methodologies must be extended outside of the procuring organization. Lean procurement systems also utilize multiple criteria [4, 10, 15] for evaluating and selecting suppliers, contradicting old purchasing systems where the price was the main concern. According to Macbeth (1987), lean procurement is now still based on reducing total costs. This implies that implementing leanness in procurement can derive savings as a large percentage of operational costs are associated with the purchase of needed materials [10]. If this is the case, then all companies should be thoroughly implementing lean in their procurement activities. However, lean procurement is not universally practiced in the industry [15], especially in the context of small and medium-sized enterprises (SMEs). The research further strengthens this by stating that lean procurement benefits are just as important to smaller companies as large companies [10].

Another gap that can be pointed out from the studies mentioned thus far is that they have not incorporated the concept of resiliency in lean procurement practices. Resiliency discussed further in detail, is an essential factor in keeping lean procurement systems adaptable and flexible to uncertainties. However, [10]. Nyazkagwa and Muthoni thoroughly discussed the lean house concept in the framework of procurement while Wilson and Roy developed a method to capture the tradeoffs of order size, frequency, and cost through a freight consolidation model [15, 16].

3 The Proposed Solution

3.1 Conceptual Model of Agents and System Elements

This research analyzes a specific case study: "The COVID-19 Disruption Treatment for Ventilators Cloud manufacturing Network". The study focuses on the impact of the 2020 pandemic on the manufacturing industry [17, 18]. The ventilator systems are utilized in situations where patients require help due to breathing problems. However, the production capacities have been compromised with the start of the pandemic and the current processes require resiliency solutions. The CM ecosystem uses absorptive and adaptive capacities to complete the production process of ventilators by using a variety of suppliers that can perform different kinds of services.

Absorptive capacities refer to the ability of a system to use its resources to return to equilibrium [19]. In the case study, several suppliers contribute to absorbing the magnitude of the disturbance before the ventilator industry changes its structure. The first stakeholder is the Ministry of Health and Medical Education (MOHME) which facilitates regulations and collaboration portals. This executive organization also provides tools and devices for health services and staffing for teaching medical knowledge. The second suppliers in the absorptive category are the hospitals. They manage several health services such as capabilities and knowledge, provision of medical inventory, and teaching staffing. Supplier number three are the ventilator companies that are used for their knowledge, experience, and intellectual property rights. Additionally, they contribute production lines for every required component, smart maintenance, and setting and training of the ventilator use. The next stakeholders are automotive manufacturers and military-oriented companies. These suppliers help with the purchasing of needed parts and providing assembly lines for the ventilators. Besides, they impart professional knowledge from highly capable experts.

The increase of specific demands in the industry can overcome the capabilities of the current and absorptive capacities. Large-scale investments in structural design are not feasible for executives. Therefore, they opt for alternative solutions to cope with the fluctuating demand. The best alternative is to use adaptive capacities. The ability to resist collapse and/or to change to sudden disruptions is described as an adaptive system [19]. Incremental adjustments allow the ecosystem to stable under larger disturbances through the addition of redundancies.

The first adaptive suppliers are medical companies. They help with health services such as importing tools and devices, and they also provide capabilities of professional knowledge. On the other hand, they can support the ventilator industry by the provision of all the required parts and with assembly lines in their plants. An important stakeholder

is the Technical Universities. Academics can support utilizing teaching; professional knowledge and they provide the necessary training. Universities, SMEs, and the owner of 3D printers can arrange the required software and devices for the manufacturing of valves and common parts. Other adaptive suppliers are Logistics companies and startups that support transportation services through the provision of vehicles.

The required services of the ventilator ecosystem and the interaction with the different suppliers are shown in Fig. 1.

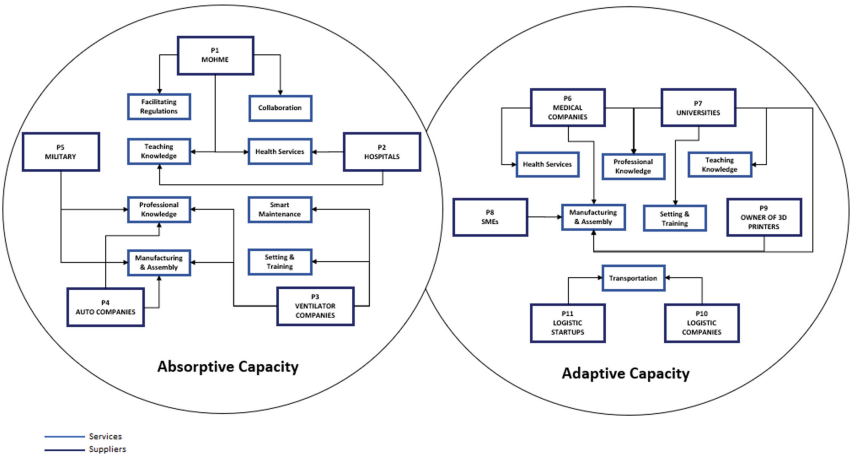


Fig. 1. Conceptual model of suppliers and services.

3.2 Proposed Idea

The capability to manage the disruptive effect of COVID-19 in terms of meeting the increasing demands for ventilators is important for all the stakeholders involved. Since the existing manufacturing capacities are not adequate to meet the currently required volumes, the existing capabilities must be improved through the cooperation of other companies in designated industries. However, it is important that while being able to meet the demand is of high importance, other organizational factors such as leanness and resilience must be considered and a model to determine the best combination of suppliers must be established.

The analytic hierarchy process (AHP) is proposed for establishing a scoring model that will help to determine the best or optimized selection of combined suppliers, referring to the services that the suppliers offer. AHP is one of the most widely used multi-criterion decision-making (MCDM) methods in the areas of planning, alternative selection, allocation, and conflict resolution [20]. It is ideal for use in complex environments that require analysis of multiple variables or criteria [21]. The method initially structures a problem into a hierarchy [22] and then conducts a pairwise comparison before determining the consistency and a sensibility analysis [23]. It translates criteria for prioritization and selection into numeric values [21, 22] and combines these criteria to

provide a compromised solution [23]. This provides a consensus for decision-making rather than just getting one definitive answer [23]. The paper will proceed to use AHP in the selection of the best suppliers by initially creating a hierarchy of the problem. Then a pairwise comparison is conducted before relative weights, the consistency ratio, and a synthesis of the results are made.

3.3 Methodology for Implementation Phase

This paper proceeds with the use of AHP as follows:

1. *Establishing the hierarchy:* Creating a hierarchy of the problem is the foundation of AHP. Based on the case study, a hierarchical structure is created starting from the main goal which then subdivides into the lower-level decision criteria. Since the suppliers offer different services, a hierarchy is established specifically to the suppliers.
2. *Pairwise comparison:* The different criteria are compared in pairs using Saaty's nine-point scale. The nine-point scale helps to identify the relative importance of one alternative from the other. From there, a comparison matrix is created. To be able to interpret and provide weights to each of the criteria, it is important to normalize the comparison matrix. The normalization is conducted by dividing each value by the total column value.
3. *Determining relative weights:* The contribution of each criterion is determined through the Eigenvector, which is also known as the priority vector. The Eigenvector is used to produce weights for each corresponding criterion using the mathematical average of all criteria. It must be noted that the sum of the values of all vectors is equal to 1.

$$Eigenvector = \frac{\Sigma \text{ normalized values in each column}}{\text{no. of weights in each column}} \times 100 \tag{1}$$

4. *Consistency ratio:* The next step is to measure the consistency index (CI). The consistency index determines the degree of the reliability and consistency of judgments. It is based on the maximum Eigenvalue which is calculated through the formula presented below.

$$CI = \frac{\lambda \max - n}{\mathrm{n} - 1} \tag{2}$$

Then the consistency ratio (CR), the ratio between the CI and the random consistency index (RI), is determined. The RI are fixed values that depend entirely on the number of evaluated criteria.

4 Case Study

4.1 Hierarchy of the Problem

The goal of the paper is the development of a resilient lean ventilator supply chain to cater to the demand of ventilators during the pandemic COVID-19. The hierarchy of the

problem consists of three different AHP Procedures with their respective main and sub-criteria, classified based on the services offered by the individual supplier alternatives are shown in Fig. 2 and detailed as:

AHP 1: 2 Criteria with 2 Sub-criteria Each. Two suppliers are evaluated under the hierarchy with a condition that at least one of them must be selected for the proposed supply chain network based on the assumptions made.

AHP 2: 2 Criteria with 4 and 5 Sub-criteria, Respectively. Seven suppliers are evaluated under the hierarchy. A combination of suppliers with the best combination of lean and resilient characteristics will be selected.

AHP 3: 2 Evaluation of Suppliers. Suppliers are evaluated under one criterion with a condition that one of them must be selected to form the supply chain network.

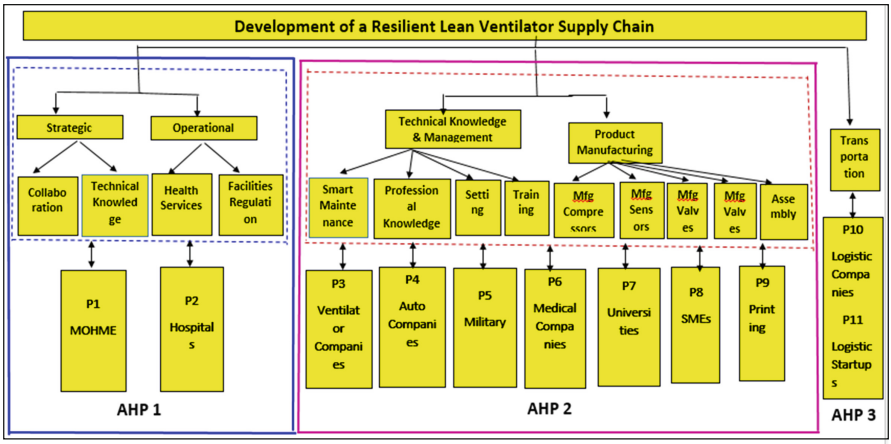


Fig. 2. Decomposition of the problem into hierarchy [24].

AHP is employed to determine the overall individual scores of the suppliers evaluated under the procedure. These scores are utilized to determine the most optimal solution from the 20 different alternative combinations of suppliers developed based on the assumptions made.

4.2 Selection of the Best Alternative

Once the individual supplier scores are calculated, the overall scores of the individual alternatives with a different combination of suppliers are determined. The alternative with the highest score is the optimal solution required. The framework with the scores for 20 selected alternatives is provided in Table 1.

Table 1. Summary of the alternatives and their rank

Alternative	Supplier combination	Overall score	Rank
Alternative 1	P1, P3, P4, P9, P10	2.056851341	1
Alternative 5	P1, P3, P5, P10	2.022287512	2
Alternative 9	P1, P3, P6, P10	1.982612511	3
Alternative 13	P2, P3, P4, P9, P11	1.868966415	4
Alternative 17	P1, P5, P6, P7, P10	1.848439264	5
Alternative 3	P2, P3, P4, P9, P10	1.600601341	6
Alternative 7	P2, P3, P5, P10	1.566037512	7
Alternative 2	P1, P3, P4, P9, P11	1.556851341	8
Alternative 11	P2, P3, P6, P10	1.526362511	9
Alternative 6	P1, P3, P5, P11	1.522287512	10
Alternative 10	P1, P3, P6, P11	1.482612511	11
Alternative 15	P2, P4, P5, P7, P10	1.412716415	12
Alternative 19	P2, P5, P6, P7, P10	1.392189264	13
Alternative 14	P1, P4, P5, P7, P11	1.368966415	14
Alternative 18	P1, P5, P6, P7, P11	1.348439264	15
Alternative 4	P2, P3, P4, P9, P10	1.100601341	16
Alternative 8	P2, P3, P5, P11	1.066037512	17
Alternative 12	P2, P3, P6, P11	1.026362511	18
Alternative 16	P2, P4, P5, P7, P11	0.912716415	19
Alternative 20	P2, P5, P6, P7, P11	0.892189264	20

Hence for the assumed supply chain requirements Option 1 is the best alternative. It consists of Suppliers P1, P3, P4, P9, and P10. In this Alternative

1. Supplier P3 provides the essential Training and Setting services
2. P4 provides spare capacity for Common Parts Manufacturing and Assembly, consistent with assumption 3
3. P9 provides the Spare capacity for Valve Manufacturing, consistent with assumption 3.

The suppliers are evaluated with adaptive capacities provided twice the weightage as absorptive capacities, implying that the suppliers are selected based on their lean performance. Hence AHP has been capitalized to arrive at a resilient-lean ventilator supply network. It is also worth noting that during the development of suppliers, P1, P3, and P10 have to be given the topmost priority and scheduled in the initial phase of the project time-plan. This is because if the implementation of the optimal solution becomes infeasible, then the project plan can be switched to the next best alternative in

the rank (Alternative 5) with maximum flexibility. This visibility is provided by AHP Process because it ranks all the other alternatives in the immediate decreasing order of preference.

5 Conclusion

This paper explores the integration of lean and resilience principles across the supply chain. The existing literature demonstrates the selection of either resilient or lean suppliers but not the trade-off between them. This research analyzes a case study about the impact of COVID-19 in the ventilator manufacturing industry and addresses this gap, contributing to the creation of a model that determines the best combination of suppliers considering the leanness and resilience factors. For this purpose, the AHP method was applied. The overall framework involves the deployment of three AHP procedures for the evaluation of suppliers across services. The framework enables the selection of suppliers to form a resilient-lean ventilator supply chain, advocating for the classification of suppliers and services into clusters based on their similarities and deducing all the possible alternatives before evaluating them against the desired conditions. This enables a decision support system to prioritize the available alternatives with AHP. These alternatives can be viewed as contingency plans. The framework hence enables the supply chain planners to develop time plans prioritizing the common steps between the best alternatives such that when the best alternative fails, the implementation can be flexibly steered to the next best option. Although the proposed method provides a systematic procedure for supplier evaluation, it has its limitations. Firstly, data collection for the AHP process is challenging. Secondly, from the perspective of supplier evaluation and the list of criteria, there would possible extensions to include several other criteria like cost, delivery, and Quality.

References

1. Handfield, R.B., Graham, G., Burns, L.: Corona virus, tariffs, trade wars and supply chain evolutionary design. Int. J. Oper. Prod. Manag. (2020). https://doi.org/10.1108/IJOPM-03-2020-0171
2. Maslarić, M., Nikolicic, S., Maslaric, M., Mircetic, D.: Digitalization of Supply Chain View project assessing the trade-off between lean and resilience through supply chain risk management Todor Backalic. Int. J. Ind. Eng. Manag. **4**, 229–236 (2013)
3. Carvalho, H., Duarte, S., Machado, V.C.: Lean, agile, resilient and green: divergencies and synergies. Int. J. Lean Six Sigma **2**, 151–179 (2011)
4. Azadegan, A., Patel, P.C., Zangoueinezhad, A., Linderman, K.: The effect of environmental complexity and environmental dynamism on lean practices. J. Oper. Manag. **31**, 193–212 (2013)
5. Birkie, S.E.: Operational resilience and lean: In search of synergies and trade-offs. J. Manuf. Technol. Manag. **27**, 185–207 (2016)
6. Yadav, V., Jain, R., Mittal, M.L., Panwar, A., Lyons, A.C.: The propagation of lean thinking in SMEs. Prod. Plan. Control **30**, 854–865 (2019)
7. Alkhoraif, A., Rashid, H., McLaughlin, P.: Lean implementation in small and medium enterprises: Literature review. Oper. Res. Perspect. **6**, 100089 (2019)

8. Wishart, M.: Business resilience in an SME context: a literature review (2018)
9. Valilai, O.F., Sodachi, M.: Inspiration of Industry 4.0 to enable a proactive sustainability assessment model through the supply chain. Procedia Manuf. **52**, 356–362 (2020)
10. Hagström, M., Wollner Tutor, M., Borgström Jönköping, B.: Implementing lean procurement opportunities, methods and hinders for medium sized enterprises-a case study. In: Implementing Lean Procurement: Opportunities, Methods and Hinders for Medium Sized Enterprises-A Case Study (2011)
11. Ali, I.: Supply chain resilience research: reviews, trends and opportunities for the future research. In: A Resilience Model for Cold Chain Logistics of Perishable Products View Project Impact of Climatic Risks on SMEs of Perishable Food Supply Chains: Moderating Role (2017)
12. Poiger, M.: Improving performance of supply chain processes by reducing variability. Vienna University of Economics and Business (2010)
13. Pereira, C.R., Christopher, M., Lago Da Silva, A.: Achieving supply chain resilience: the role of procurement. Supply Chain Manag. **19**, 626–642 (2014)
14. Bauer, D., Göbl, M.: Flexibility measurement issues in supply chain management. J. Appl. Leadership Manage. **5**, 1–14 (2017)
15. Wilson, M.M.J., Roy, R.N.: Enabling lean procurement: a consolidation model for small-and medium-sized enterprises. J. Manuf. Technol. Manag. **20**, 817–833 (2009)
16. Nyakagwa, G.O., Muthoni, D.K.: Factors affecting implementation of lean procurement in multinational enterprises: a vase study of British AmericanTobacco (KENYA). Int. J. Soc. Sci. Entrep. **1**, 395 (2014)
17. Rezapour Niari, M., Eshgi, K., Fatahi Valilai, O.: Topology analysis of manufacturing service supply–demand hyper-network considering QoS properties in the cloud manufacturing system. Robot. Comput. Integr. Manuf. **72**, 102205 (2021)
18. Rezapour Niari, M., Eshghi, K., Fatahi Valilai, O.: Adaptive capacity management in cloud manufacturing hyper-network platform: case of COVID-19 equipment production. Int. J. Manag. Sci. Eng. Manag. 1–20 (2022). https://doi.org/10.1080/17509653.2021.2009389
19. Tanner, T., Bahadur, A., Moench, M.: Challenges for resilience policy and practice. ODI Working Paper 519 (2017)
20. Tramarico, C.L., Mizuno, D., Salomon, V.A.P., Marins, F.A.S.: Analytic hierarchy process and supply chain management: a bibliometric study. Procedia Comput. Sci. **55**, 441–450 (2015)
21. Badea, A., Prostean, G., Goncalves, G., Allaoui, H.: Assessing risk factors in collaborative supply chain with the analytic hierarchy process (AHP). Procedia - Soc. Behav. Sci. **124**, 114–123 (2014)
22. Waris, M., et al.: An application of analytic hierarchy process (AHP) for sustainable procurement of construction equipment: multicriteria-based decision framework for Malaysia. Math. Probl. Eng. **2019**, 1–20 (2019)
23. Salomon, V.A.P., Tramarico, C.L., Marins, F.A.S.: Analytic hierarchy process applied to supply chain management. In: Applications and Theory of Analytic Hierarchy Process - Decision Making for Strategic Decisions (2016). https://doi.org/10.5772/64022
24. Rouyendegh, B.D.: Selecting the best supplier using analytic hierarchy process (AHP) method. African J. Bus. Manag. **6**, 1455–1462 (2012)

Fighting Hospital Infections with Engineering 4.0

Roberto Mosca, Marco Mosca(✉), Roberto Revetria, Fabio Currò, and Federico Briatore

Mechanical, Industrial and Transport Engineer Department (D.I.M.E.), University of Genoa, 16126 Genoa, Ge, Italy
mosca@diptem.unige.it, marcotulliomosca@gmail.com, roberto.revetria@unige.it, 4452164@studenti.unige.it

Abstract. In the repeated interventions carried out by the authors in the healthcare sector [1–4] (hospitals, outpatient clinics and clinics), including assistance facilities (residences for the elderly and outpatient medical offices) the problem of so-called hospital or nosocomial infections has always been reported to the team by the medical and nursing staff. Starting from an age-old experience of sanitization of confined environments, achieved by the authors by using a 4.0 machine, for the production of gaseous ozone and UVC rays [8], it was required to the team to extend the benefits achieved to the healthcare sector. This goal was possible by generating a dedicated approach, for an effective action to combat this serious problem of global significance. The machine mentioned was conceptualized, designed and developed by the authors by specific Engineering 4.0 methodologies, meaning with this term the use of all Engineering technologies, techniques, software, tools, and devices characterizing the fourth industrial revolution.

Keywords: Hospital infections · Sanitization 4.0 · Ozone · UVC

1 Identification of Risk Phases

In the normal management of an operating block, both patients and operators face continuous risks for their safety. It is known, in fact, that the surgical patients are, not infrequently, subject to aggression by viruses and bacteria, to heal which it is necessary to intervene with difficult explicit treatments, due to the resistance developed by pathogens to drugs, in operating blocks. The direct consequence is the infections for the patients and the related costs for the institution, which must provide appropriate treatment. In addition to this, there is the possible presence of insects, molds and even rodents. To combat these entities is not trivial due to the impossibility of intervention with common insecticides and rat poisoning, in a sterile environment. Another problem frequently encountered, especially in periods of overload of interventions, is the fact that a surgical instrument escapes the control of the operators being left into the patient's body, with the need to provide for its urgent removal. In this case it is easy to imagine the impact on the patient and for the institution. Then, with regard to the staff responsible of washing the surgical irons arriving "dirty" from the operating rooms requiring a

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 310–317, 2023.
https://doi.org/10.1007/978-3-031-16281-7_30

pre-treatment upstream to the sterilization cycle, it must be considered the risk of cuts or punctures, with easily imaginable consequences.

2 Hospital Infections

From the literature of the sector, it is highlighted what is reported in the following regarding this important problem [10]. Hospital or nosocomial infections are infectious diseases acquired in hospitals or in healthcare settings (like nursing homes, long-term hospitalizations, long-term clinics, residences for the elderly, etc.) that affect between 5% and 10% of patients, causing 80,000 deaths a year in the US alone. To be defined as hospital infections, the patient must have been hospitalized for a cause other than the infection in question and must not have signs of an infectious disease being incubated at the time of admission. It is therefore necessary to be sure that it was contracted in the facility. Such infections normally occur 48 h after hospitalization, or up to 3 days after dismission, or up to 30 days after an operation. The timing varies according to the type of infection and the viral load affecting the patient. The most frequent are septicemia due to venous access, where a needle has been inserted (about 50% of cases of infection), pneumonia and infections of the respiratory tract (21%, constantly increasing), urinary tract (10.5%). According to more recent studies, today they amount to 30/40% of the total hospital infections, attacking the skin and soft tissues (9.8%), the ear, nose, throat and eye (3%), the osteoarticular system (2%), the nervous system (1%), the cardiovascular system (0.3%), the surgical and decubitus wounds (8% of hospitalized, and between 15 and 25% of hospitalized wards for long hospitalization). The interventions, on the other hand, can cause a wide series of infections, ranging from the superficial ones of the skin to the very deep ones of tissues, organs and implants. Colon surgery in Italy is a highly risky procedure, with an infection rate of almost 9% of interventions. A serious problem is that 70% of the bacteria involved are resistant to common antibiotics as they are used to enduring very strong pharmacological pressure. The simple washing of hands before any operation leads to the prevention of 25% of infections, therefore a system of sterilization and sanitization of the environments and objects touched by the operators leads to a strong reduction in the chances of infection. Pathogens are various and the predominant ones in Italy are staphylococcus aureus (36% of cases in the south, 34% in the north and 25% in the center), Pseudomonas aeruginosa (32% of cases in the center, 30% in the south and 19% in the north), escherichia coli (25% of cases in the north, 18% in the center, 13% in the south), staphylococcus epidermidis (12% of cases in the center, 10% in both north and south), enterococcus faecalis and enterococcus fecium (7% of cases in the center, 6% in the north, 5% in the south), klebsiella pneumoniae (6% of cases both in the south and in the center and in the north).

3 Sanitization 4.0

With regard to the problem of sanitization, which is a crucial in the management of operating blocks, important improvements can be brought to traditional methods, using engineering 4.0 and its technologies for production and control of UVC rays and gaseous

ozone. A relevant aspect in an operating block 4.0 is guaranteeing a constant sanitization, by means of UVC rays, of the "filter area", which is the room located in front of any operating room. Also, to protect the operators, who not infrequently get injured by treating dirty surgical instruments (major concern), it was developed a "hand free" pre-sanitization method of the surgical equipment trays. It is performed by a machine of under pressure pre-washing and UVC treatment, by appropriate 4.0 technology, conceptualized and developed by the authors. The proposed process of environmental sanitization consists in 2 steps: a) quick treatment, the operating block is treated, with UVC rays, controlled in wavelength and frequency, for only a few minutes between one patient and the next, providing major guarantees from the current sanitization; b) deep treatment, at the end of the shift, a radical treatment is provided with gaseous ozone, produced centrally by a special 4.0 plant, thus guaranteeing deep sanitization against viruses, bacteria, insects, fungi, spores and so on (because capable of penetrating in fabrics and mattresses, as scientifically recognized by the literature of sector on a planetary level). This sanitization performance is achieved with relevant advantages like quick treatment, at negligible costs, and with a fast resume of the room to re populate. In fact, both the proposed treatments do not leave any residue in the environment. Treatments are conducted in the absence of personnel, not generating risks or requiring subsequent cleaning. For an in-depth analysis of the subject in question, please refer to the full text of the paper presented by Marco Mosca at the IAENG-WCE 2021 (International Association of Engineers, Word Congress on Engineering), publication awarded as best paper for the bioengineering section.

In short, the method proposed for the sanitization of environments consist 2 phases, by means of the machine designed and developed by the authors, described in paper [8]:

- irradiation with UVC rays (at a fixed distance and for tabulated timing);
- exposure to a certain concentration of ozone (produced from the air, at a negligible cost) for tabulated timing, according to the agent to be sanitized. The machine is capable to reproduce the Chapman-cycle in the environment, as it occurs in the Ozonosphere. As the wavelength of UVC rays varies, the oxygen present in the environment follows the transformations:
- under certain wavelengths (100–240 nm) it splits the O_2 into $O + O$ which, combining with the other O_2 present, form O_3;
- under other wavelengths (240–315 nm) the O_3 are transformed back into O_2 and O, restoring the initial equilibrium.

The 4.0 machine detects, by means of a remote sensor (positioned in the most distant corner of the room from the machine), the concentration of ozone, continuously modulating the production of the gas in consideration of the volume of the room, the environmental geometry and the possible dispersions. This maintains the concentration of ozone for the time necessary to the sanitization. After that, the machine completes the cycle restoring the livability conditions of the treated rooms. The technologies proposed (UVC and ozone), of equal effectiveness, have different characteristics and applications: UVC rays act by radiation in proximity, in a few minutes (they do not cross surfaces, mattresses and sofas). They are suitable for quick sanitization between the use of one

patient and the next. Gaseous ozone takes a bit more time (according to the concentrations required and tabulated timing), but has the ability to sanitize the entire environment, not stopping on the surfaces and deeply penetrating fabrics and mattresses. It is therefore suitable for a thorough sanitization at the end of the shift.

Fig. 1. UVC lamps

The main strengths consist in the possibility of dry sanitization without leaving residues (it does not require cleaning after treatment), free of harmful and permanent substances in the environment, continuous (part-time use of an operator assigned to other tasks), compatible with desks and bookcases (does not damage electronics, paper and furnishings), fast, economical (costs less than vaporized substances, which are a daily expense and costs less than sanitization performed by external companies, being designed for energy saving (low consumption starting from 14W), and it is a safe sanitization. In synthesis, to reach the goal of a sanitation capable of effectively and efficiently combating the problem of hospital infections, the authors identified the need to proceed with a capillary and systematic intervention, fully covering each area of the operating block. For this reason, the block has been divided in 5 areas: 1) access, 2) filter rooms, 3) operating rooms, 4) sterilization room, 5) air conditioning system. Optimal sanitization processes were studied for each area in consideration of the specific needs, such as illustrated in the following. Please see references number [9–27].

3.1 Sanitization of Access to the Operating Block

People (medical and nursing team, operators, cleaning personnel, maintenance workers and patients) and things (medical aids, drugs, machines, products,…) necessarily have almost continuous access to the operating block. The block requires higher hygienic conditions than the rest of the structure and must be protected from pathogens, viruses, bacteria, molds, fungi, spores, insects and even rodents. Consequently, the access to the block is configured as a critical gateway to the hygienic preservation and, therefore, it is vital to ensure the sanitization of each entering element. For this purpose, action is taken both on people (by means of the restricted transit on a sanitizing platform with emission of UVC rays) and on things (by conveying them through a tunnel, in which they slide on a belt while irradiated with UVC rays).

3.2 Sanitization of the "Filter Room"

Access to the operating rooms is possible only through a small room in front, called filter, characterized by two sliding doors (positioned upstream and downstream of the filter) with the purpose of separating the operating room from any upstream process step. In the filter, in fact, the staff abandons clothes and footwear from outside and wears clean clothing. The identified risk obviously consists in cross-contamination during the change of clothes and shoes which, contaminating the floor of the filter, causes the clean shoes worn by the staff to lose the sterility. Same for the clothes, stored in lockers, which gradually get dirty. For this purpose, it is proposed to intervene on the filter room with an almost continuous sanitization process, by means of a double sanitization intervention: during normal activity by means of UVC rays (kept in operation after of each transit) and, after the end of each shift, by using gaseous ozone.

3.3 Sanitization of Operating Rooms

In literature there are many examples of infections, even serious ones, contracted by patients in the operating rooms. With reference to the chapter previously treated on hospital infections, unfortunately, the event is not uncommon. It shall be considered, as reported, that hospital infections, due to continuous pharmacological attempts to counter them, are particularly persistent, dangerous and difficult to heal. For this purpose, it is proposed to intervene on the surgery with an almost continuous sanitization process through the use, appropriately combined, of UVC rays (quick sanitization at each use) and ozone (complete sanitization at the end of each shift).

3.4 Sanitization of the Sterilization Room

The sterilization process of the surgical instruments leaving the operating rooms consists mainly of two phases: washing and sterilization. The washing phase involves the "dirty" area of sterilization in which the operator performs the following activities: 1) transfer of the surgical instruments from the tray to the washing tank; 2) soaking in sanitizing liquid; 3) rinsing. The sterilization phase involves the "clean" area of the sterilization in which the operator performs the following activities: 1) wrapping and sealing the surgical instruments; 2) autoclaving; 3) deposition of wrapped surgical instruments. The sterilization room, in the dirty area, presents an important risk phase for the operator who, statistically, is subject to the likelihood of being punctured and cut with the surgical instruments, during the washing phase. This means, for the operator, exposure to the risk of contracting serious diseases such as, HIV and Hepatitis. For this purpose, it is proposed to intervene on the sterilization room with a "hand free" sanitizing machine specially engineered for the purpose, to be placed upstream the washing phase, to sanitize the surgical instruments before the contact with the operator. The characteristic phases of the sanitization process operated by the machine are: a) loading of the tray coming from the operating room; b) rinsing the surgical instruments without having to touch them (the machine receives the tray directly); c) high pressure pre-wash; d) UVC treatment. It shall be specified that the 4.0 machine does not replace the current sterilization procedures but is integrated in current process.

3.5 Sanitization of the Air Conditioning System

The air conditioning system generates, notoriously, a risk phase for the spread of pathogens from one room to another. Viruses, bacteria, molds, fungi, spores, insects and rodents tend to lurk in the ventilation ducts and, part of them, in the filters. The cleaning of the filters, in turn, presents other risk phases such as incorrect timing of sanitization, incorrect sanitization and the operator contact with contaminated filters. According to a basic principle of resilience, it is proposed to exploit this weakness as a critical factor for success, e.g., to use UVC rays to keep the pipelines sanitized, thus drastically reducing the opportunity for contagion from room to room and, not less important, using the ducts as a vehicle for gaseous ozone, during the rest shifts, to sanitize the entire operating block (no bulky machines to move among the room). For this purpose, it is proposed to intervene on the rooms with a sanitization system based on the double technology previously described. Nevertheless, the configuration for this specific application is different: UVC rays are installed inside the ducts (cycle operation), while the ozone machine is installed upstream of the ducts, in this way by exploiting the ducts for the distribution of ozone in the rooms (operation at the end of the shift, with rooms evacuated). The ozone produced will be disposed-off by reversing the Chapman-cycle, before re-populating the rooms.

4 Technical and Methodological Details

A UVC-Ozone machine equipped with 12 lamps of 14 W/each is able to sanitize an environment of 100 m^3 in 20 min, conducted by an operator who can simultaneously manage 2 machines used in contiguous rooms. The activities carried out are positioning, switching on and shutdown. These tasks are normally assigned to the cleaning personnel, as part of normal activity (as written, 20 min at the end of each shift).

5 Economic Analysis (on a Case Study)

An operating block on 4 floors of 250 m^2/each is considered. Each floor is divided into 15 rooms, 5 bathrooms and an access corridor. For the sanitization of each floor, 2 machines costing € 4,000/each will be used, with a total investment of € 32,000. The pure cost of sanitizing the entire block, as seen in Fig. 2, is € 27.

Considering a payback time of one year, the cost of an intervention is € 61, because 2 shifts/day are performed for 15 days/month, with a total of 360 sanitizations/year.

The impact of the investment on each sanitization in the first year is 12,000€/360 d = 34€, so the cost of a sanitization of the entire structure in the first year is equal to 27€ + 34€ = 61€, a cost that from the second year drops to 27€. This figure should be compared with the current € 1,500/sanitization paid to external companies. The case study demonstrates the indisputable advantages of the proposed system compared to traditional technologies: absolute sanitization in a short time, negligible costs, no use of chemicals, environments that can be readily re-inhabited at the end of the treatment. The machines require, as a single maintenance intervention, the replacement of the lamps and minor accessories every 12,000 h of operation, with a cost of maintenance of €

Cost per treatment (full building)			
Electricity consumption per lamp	14	[Wh]	B
Llights on per machine (at the same time)	12	[#]	C
Electricity consumption per machine	168	[Wh]	D=BxC
Electricity consumption per machine	0,168	[kWh]	E=D/1.000
Cost per kWh	0,2	[€]	F
Operating time (2 machines)	20,0	[h]	A
Electricity cost per treatment	0,67	[€]	G=ExFxA
Machine directly operating time	10,0	[h]	K
90%, Efficiency saving (along treatment)	-9,0	[h]	H= -Kx90%
4%, Operator extra time (move, setup)	0,4	[h]	I= Kx10%
Operating time (1 man + 1 machine)	1,4	[h]	L=K+H+I
Manpower general cost (full cost)	16	[€/h]	M
Manpower cost per treatment	22,5	[€]	N=LxM
Maintenance cost (8 machines for 6Kh)	4,0	[€]	J
Total cost per treatment	**27**	**[€]**	O=G+N+J

Fig. 2. Cost per treatment

9,600 (total cost for 8 machines, 2 per floor). The useful life of a machine is estimated at 20 years. The incidence of maintenance, in this case study, amounts at 4€/treatment, as it can be seen in Fig. 2.

6 Conclusions

Thanks to what has been made available by Engineering 4.0, the authors conceptualized, developed and created a range of 4.0 machines that, used according to the approach proposed, allow to effectively combat the devastating hospital infections that today are causing many deaths of patients in transit through operating blocks. This was the first step towards a 4.0 concept of the entire operating block, safer for staff and patients. The washing and sanitizing machine described in paragraph 3.4 is another intervention carried out by the authors in this direction. They are working intensively on this and soon the operating block 4.0 will be available for an industrial level realization. At this point, the authors will have achieved the goal they set when they dedicated to this study, which is to make a contribution to medical colleagues, to alleviate the suffering to which our humanity is subjected every day.

References

1. Cassettari, L., Mosca, M., Mosca, R., Rolando, F.: An healthcare process reengineering using discrete event simulation, In: Proceedings of the World Congress on Engineering and Computer Science 2013 Vol II WCECS 2013, 23–25 October, San Francisco, USA, Lecture Notes in Engineering and Computer Science.. **2**, pp. 1174–1179 (2013)
2. Cassettari, L., Mosca, R., Morrison, B., Rolando, F., Revetria, R., Orfeo, A.: A system dynamics study of an emergency department impact on the management of Hospitals surgery activity. In: Proceedings of International Multiconference WCE 2015 (2013)
3. Bendato, I., Cassettari, L., Mosca, R.: Improving the efficiency of a hospital emergency department. In: SOMET 2015 International Conference (2015)

4. Patrone, C., Cassettari, L., Mosca, R., Damiani, L., Revetria, R.: Optimization of lean surgical route through POCT acquisition. In: International multiconference of Engineers, IMECS (2017)
5. Cutler, T.D., Zimmerman, J.J.: Ultraviolet irradiation and the mechanisms underlying its inactivation of infectious agents. Anim. Health Res. Rev. **12**, 15–23 (2011)
6. Tseng, C.C., Li, C.S.: Inactivation of viruses on surfaces by ultraviolet germicidal irradiation. J. Occup. Environ. Hyg. **4**(6), 400–405 (2007)
7. Rabenau, H.F., Kampf, G., Cinatl, J., Doerr, H.W.: Efficacy of various disinfectants against SARS coronavirus. J. Hosp. Infect. **61**(2) pp. 107–111 (2005)
8. Mosca, R., Mosca, M., Revetria, R., Cassettari, L., Currò, F., Galli, G.: Sanitizing of confined spaces using gaseous ozone produced by 4.0 Machines. In: WCE ICSBB_210312Rx, accepted and included in the conference proceedings published by IAENG (ISBN: 978–988–14049–2–3) – BEST PAPER AWARD (2021)
9. Healthcare-associated infections in intensive care units (Annual Epidemiological Report for 2015)
10. Italian Ministry of Health, protocol 24482 of 31 Aug 96
11. Muzhi, Z.: China.org.cn 26 Feb (2020). http://www.china.org.cn/opinion/2020-02/26/content_75747237.htm
12. Martinez-Sanchez, G., Schwartz, A., Di Donna, V.: Potential cytoprotective activity of ozone therapy in SARS-CoV-2/COVID-19. In: Antioxidants (2020)
13. Castaño, N., et al.: Fomite transmission and disinfection strategies for SARS-CoV-2 and related viruses, Arxiv.org (2020)
14. Tseng, C., Li, C.: Inactivation of surface viruses by gaseous ozone. J. Environ. Health **70**, 56–62 (2008)
15. Cristiano, L.: Could ozone be an effective disinfection measure against the novel coronavirus (SARS-CoV-2). J. Prev. Med. Hyg. **61**, 301–303 (2020)
16. Zhou, M.: China.org.cn (2020). http://www.china.org.cn/opinion/2020-02/26/content_75747237.htm
17. Shin, G.A., Sobsey, M.D.: Reduction of Norwalk virus, poliovirus 1, and bacteriophage ms2 by ozone disinfection of water. Appl. Environ. Microbiol. **69**, 3975–3978 (2003)
18. Govindaraj, S., Muthuraman, M.S.: Systematic review on sterilization methods of implants and medical devices. Int. J. ChemTech Res. **8**, 897–911 (2015)
19. Burkhart, C.G.: Ozone disinfectants like SoClean CPAP sanitizer can be used to sterilize cloth and N95 masks in the protection against COVID-19. Open Dermatol. J. **14**, 14–15 (2020)
20. Fischer, R.J., et al.: Effectiveness of N95 respirator decontamination and reuse against SARS-CoV-2 virus. Emerg. Infect. Dis. **26**, 2253–2255 (2020)
21. Rutala, W.A., Weber, D.J.: Disinfection and sterilization in health care facilities: an overview and current issues. Infect. Dis. Clin. North Am. **30**, 609–637 (2016)
22. Ebihara, K.: Sterilization characteristics of ozone-mist spray for chemical-free agriculture. Int. J. Plasma Environ. Sci. Technol. **10**, 11–15 (2016)
23. Edelstein, P., Whittaker, R.E., Krelling, R.L., Howell, C.L.: Efficacy of ozone in eradication of legionella pneumophila from hospital fixture. Appl. Environ. Microbiol. **44**, 1330–1334 (1984)
24. Joret, J.C., Block, J.C., Richard, Y.: Wastewater Disinfection: elimination of feal bacterial and eneric viruse by ozone. J. Int. Ozone Assoc. **4**, 91–99 (1984)
25. Farooq, S., Akhlaque, S.: Comparative response of mixed cultures of bacteria and virus to ozonation. Water Res. **17**, 809–812 (1983)
26. Harakeh, M.S., Butler, M.: Factors increasing the ozone inactivation of enteric viruses in effluent. J. Int. Ozone Assoc. **6**, 235–243 (1984)
27. Kawamura, K., Kaneko, M., Hirata, T., Taguchi, K.: Microbial Indicators for the Efficiency of Disinfection Processes. Water Sci. Technol. **18**, 175–184 (1984)

Cyber-Physical Equipment as a Service

Gustavo Sanchez[1(✉)], Giancarlo Bo[1], Fabrizio Cardinali[1], and Flavio Tonelli[2]

[1] MYWAI™ SRL, Via Portobello 19, Baia del Silenzio, 16039 Sestri Levante, Genoa, Italy
g.sanchez@myw.ai

[2] Department of Mechanical, Energy, Management and Transportation Engineering (DIME), Polytechnic School, University of Genoa, Genoa, Italy

Abstract. To transition from a traditional product-based to a service-based business model, manufacturing companies need a reliable, efficient, and affordable technological infrastructure. Moreover, in this paper we hypothesize that companies going through this transition, should consider themselves not only as original equipment manufacturers (OEMs), but as providers of specific Cyber Physical Equipment as a Service (CPE-aaS), following the same digital transformation pathway experienced in other markets, from services to media. First, to clarify our approach, we introduce basic concepts and terminology. Then, general aspects related to the role of edge platforms blending new generation IoT and AI advancements (AIOT), considered as a crucial enabler technology in this context, are discussed. Finally, to illustrate our approach, two successful practical implementations are briefly presented: projects RAISE and PROMENAIDE respectively, addressing Robotic and Medical Equipment servitization for two world leading companies, namely Mitsubishi Electrics™ and ESAOTE™.

Keywords: AIOT · Smart factory · Servitization · Cyber-Physical Systems · Cyber Physical Equipment · Internet of Things · Edge computing · Artificial neural networks · Explainable AI · Trustworthy AI

1 Introduction

Although the term *cybernetics* was introduced by Wiener back in 1948, to describe the interplay between control and communication systems [1], only recently have we witnessed an astonishing interest in understanding and engineering Cyber-Physical Systems (CPS), both from a theoretical and practical point of view.

In this paper, we define a CPS as an environment cohabited by a set of *cyber* components for computation/control functions $C = \{C_1, C_2, \ldots, C_{nC}\}$, a set of *physical* components for sensing/acting functions $P = \{P_1, P, \ldots, P_{nP}\}$ and a set of *communication* components (links) $L = \{L_1, L_2, \ldots, L_{nL}\}$.

Since each of its components can be described by its own set of dynamic equations, the complexity of models to accurately describe the behavior of a CPS can scale up very quickly.

Moreover, we define an Industrial IoT (IIoT) platform as a type of CPS, comprising industrial assets (equipment, machines, etc.), sensors/actuators, cloud/edge computing

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 318–327, 2023.
https://doi.org/10.1007/978-3-031-16281-7_31

components, gateways, etc. enabling analysis and advanced decision making with the final goal of optimizing the overall behavior of the production/service process.

Finally, we entrust that a new model is emerging, which we propose to refer to as CPE-aaS (Cyber-Physical Equipment as a Service) where simplified CPS advancements are adopted in the conceptual design and manufacturing of new devices, appliances and machinery (defined here in general terms as "Equipment", following the ISA S88 or IEC 61512-1 standard) with related open interoperability standards with the final goal of enabling a recurring service business model, along the lines of the "as a service" model, which has already disrupted other industries far before the manufacturing market.

Although we acknowledge that other related expressions are already commonly preferred in the literature, e.g. Industrial Product Service Systems (IPS^2), Product as a Service (PaaS), Equipment as a Service (EaaS), Cyber-Physical Production Systems (CPPS), etc. we think that adopting the concept of CPE-aaS has the advantage of highlighting the cyber-physical nature of the system whilst keeping the "as a service" expression, which has become highly recognizable nowadays.

As co-authors of this paper could recently observe during recent fairs (e.g. the 2022 SPS, Smart Production Solutions fair in Parma, Italy), an increasing number of companies are currently offering industrial products with a blend of AI and IoT capabilities, ranging from smart sensors to full-fledged digital AI platforms. We would define such trend as AIoT. In general, various smart and connectivity features, are being integrated to traditional devices/equipment with an additional AIoT layer. Only in Italy, according to the current National Recovery and Resilience Plan, the investment in projects related to digital transformation and sustainability using an advanced blend of IoT and AI advancements at their basis are estimated in 100 billion euros [2].

Another trend which is possible to observe nowadays is *servitization*, referring to the process by which a manufacturing company offers an increasing number of services, in addition to their traditional product and spare parts catalog, so that new competitive advantage and/or revenue streams are generated. Such services have started to be focused on what may be defined as "Service-tech" offering new after sales service models starting from new forms of prognostic and prescriptive maintenance, which have been easier to introduce and embrace as first enhanced services to be offered on now fully connected products. After such first wave of enhanced maintenance solutions new services are kicking in from new machine services verticals which may broadly be defined as "Ener-tech" (e.g. machine to grid energy optimization), Fin-tech (e.g. pay per use or pay per outcome of the machinery), and/or Insur-tech (e.g. pay per event, parametric insurance of production losses due to machinery halting or break down…).

As an example, consider the cycle of services showed in Fig. 1. In a simplified way, given a manufacturing company, its servitization level can be defined as the number of services it concurrently offers.

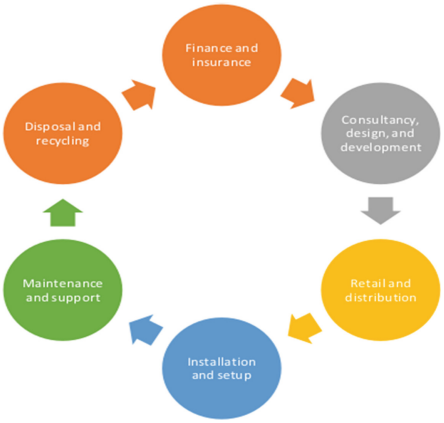


Fig. 1. Cycle of services

From data available in 2018, including 190.000 manufacturing companies in 114 countries, it was possible to observe that USA, UK and Australia were the countries with highest levels of servitization: at least 50% of manufacturing companies in those countries offered at least one service, support and maintenance being the most offered [3].

In fact, the concept of Maintenance-as-a-Service can be traced back to the Rolls Royce's "power-by-the-hour" program, usually considered as a benchmark in terms of performance-based maintenance services: repair and overhaul were offered with fixed price for each hour the engine was in the air. Note that the Rolls Royce's case can be considered as the precursor of today's subscription models. It was enabled by a combination of factors, chief among which is the development and massification of reliable digital control systems [4]. Once the required sensors, actuators, and signal processing devices were in place, the collected data became available to AI platforms to perform higher level analysis, beyond the basic control function, as depicted in Fig. 2.

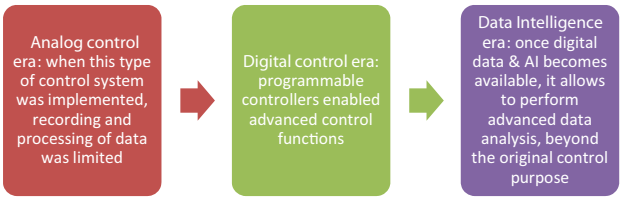


Fig. 2. Digital control as enabler of higher-level analysis

At this point it is easy to understand why the next step of industrial machinery manufacturing must be its evolution towards an Equipment-as-a-Service (EaaS) business model. In this model, not only specific functions are servitized, but ***all*** the necessary functions for the equipment operation are offered as a service, shifting in practice from

one-off sales of capital goods (CapEx) to recurring revenue streams based on usage or outcomes (OpEx).

Transitioning to an EaaS business model is a strategic decision which deserves thoughtful consideration being comparable to a Product Service System (PSS) approach for which success depends on complex dynamics [5]. It is possible to mention several arguments for and against it. One of the most important positive drivers is that nowadays equipment users are expecting flexible solutions, to be agile with their processes, to scale up and down quickly in response to market fluctuations, which is one of the important lessons from recent Covid-19 pandemic and the Russia-Ukraine conflict which dramatically influenced planetary supply and production chains in the recent past. OEMs are now in a similar situation compared to software companies since their customers expect now frequent *upgrades* to their equipment with new features and capabilities to be provided "as a service" on an ongoing basis.

On the other hand, the most important argument against, is that ownership of equipment may still be considered important from the accounting or financial point of view. Moreover, OEMs need to be prepared to observe their revenues to decline initially when they begin to offer EaaS subscriptions. However, at a subsequent time, in general this trend becomes inverted: this is known as the financial "fish" pattern, as showed in Fig. 3. Readers interested in more details about this phenomenon, as well as understanding good practices to transition to a service-based business model, can refer to [6].

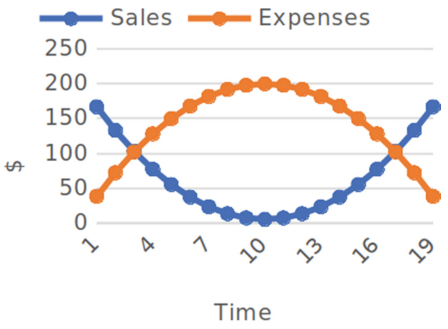


Fig. 3. The financial "fish" pattern

In the present paper, we explore the concept of CPE-aaS, together with the role of AIoT edge platforms in their practical implementation. To the best of our knowledge, there is a scarcity in the public literature concerning this topic, one of the previous works being [11], in which authors address how technologies like the Internet of Things (IoT), Cloud Computing (CC) and Predictive Analytics (PA) facilitate service transformation in industrial companies.

Specifically, we try to trigger the reflection about the following questions: *which elements should consider OEMs before taking the decision to offer CPE-aaS? How should OEMs plan this transition? What are the human and technical requirements for this transition? Which basic functions should be able to perform the AIoT platform?* The rest of the paper contains the following sections. In Sect. 2, we focus on the general relationship between EaaS and AIoT platforms, trying to answer the previous questions from

a general point of view. In Sect. 3, two successful implementation cases are presented, in which co-authors have been involved. Finally, conclusions are given.

2 Equipment-as-a-Service and IIoT

If, after careful consideration, an OEM decides to transition to EaaS model, then each step in the road must be thoroughly planned, both in terms of required people and technologies, which we briefly discuss next. In Fig. 4, we present a very simplified list of the steps that are usually necessary. Note however that it has been previously highlighted that defining an integration methodology to start offering smart services is a challenging problem, which requires the integration of components and methods from heterogeneous domains [13].

Fig. 4. Steps to transition to EaaS model

During the first step it is necessary to ask questions like:

- Which services should be offered, at what price?
- Which costs/risks must be considered?
- Which are the target customers?

Once these questions are considered, it is possible to think about technical aspects like operations and IT requirements, billing systems, connectivity infrastructure, etc. Then, usually a pilot program is run to assess the feasibility of the project, which in case of success, is to be scaled until reaching the new desired stationary state.

Again, it is important not to underestimate the complexity of this transition process. It is well-known that few companies have successfully transitioned to digital models when confronted with disruptive entrants [7]. Besides, human/managerial strengths are central for successful transformation: it is not only about hardware and software technologies. OEMs should consider adopting methodologies like *Design Thinking* to consider aspects like user experience (UX), sustainability [8], finance, social values, ethics, etc. to integrate all these dimensions into their business model. A new paradigm is required, which can be challenging for OEMs: they should start considering themselves as manufacturers of connected and smart equipment, or as mentioned in the title of this paper, they should consider themselves as providers of *cyber-physical equipment* as a service (CPE-aaS).

Equipment buyers are already looking for smart options, with advanced self-diagnostic functions, even within the scope of traditional ownership models. Therefore, a flexible platform, to expose their data securely will always be one of the most important requirements. Each equipment will now be considered as a data server, not only as a traditional asset. Moreover, they will also be expected to consume and process data smartly.

In their famous book, Baines and Lightfood (2012) state that "*successful delivery of advanced services is enabled by information and communication technologies*". Indeed, it becomes difficult nowadays to think how to deliver advanced services *without* the use of ICTs [9]. Note that once the IoT platform is installed and working on an Equipment, it should be exploited as a part of a larger ecosystem where AI may be used to address and possibly enhance what are the "vital parameters" during an equipment lifecycle: quality, maintenance, user health & safety, environment etc.

In this regard, an ideal AIoT platform should allow the following functions:

1) Collect and process data from the industrial equipment, which can be used for different applications: designing future version of the equipment, optimization of operations, triggering service activities, supporting insurance claims, billing, etc.
2) High-level remote-control functions, related to diagnostics, maintenance, software upgrading, disposal, etc.
3) Offer persistent, responsive alignment of data streaming from real equipment to digital twins to be used to possibly uniquely identify and control machinery either in the real world or its digital counterpart (what many today start calling the "productive Metaverse") where many of the simulations and verifications may happen prior to affect the real "thing".

Therefore, the information flow can go from OEM applications to equipment applications or from third party applications and equipment applications, as shown in Fig. 5.

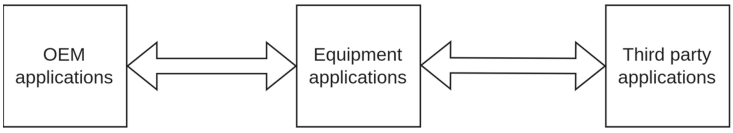


Fig. 5. Information/control flow between different applications

There are many types of AIoT platforms, devoted to PSS, suitable to different requirements. Some are specific to one sector, like healthcare, energy, or agriculture. Others are more general in scope and need to be customized. Finally, many platforms are increasingly adding the availability of specialized intelligence as their prime goal to be made available whenever and wherever the equipment needs some specific intelligence for its functioning and servicing, from the cloud to the fog gateway to the edge and as far as the "very edge" where ideally all the intelligence the CPE needs for its correct functioning, is hosted "inside" the physical boundaries of the Equipment.

In this regard, authors of this paper have been working on a platform called MYWAI™ EaaS, which can be defined as an AIoT platform optimized for delivering Equipment Intelligence whenever needed to offer better services and returns to the OEM using it by integrating state of the art IoT, AI and Distributed Ledger Technology (DLT) components. It is an open, flexible, modular platform ready to deliver intelligence to machinery, wherever it is required, running state-of-the-art machine learning algorithms

with the final goal of helping OEMs to meet their customer's expectations with new servicing and business models, based on the following principles:

1) Focus on OEMs' specific needs/expectations: although there are many pre-installed features, the platform can be fully customized.
2) Security, safety, and reliability: based on DLT.
3) Full support for clients during their "servitization" journey.
4) Seamlessly and swift integration through most common IIoT protocols.
5) Explainability and trustworthiness of all deployed AI intelligence.

3 Experimental Use Cases

The concepts behind the platform mentioned in the previous section have been developed and tested in the framework of important EU and Italian RD projects, having as a reference real life use cases and practical requirements, constraints, and expectations, in terms of production performance and life cycle optimization. Two relevant examples of such achievements are briefly introduced in the following paragraphs.

The *RAISE*™ - *Robots as an Intelligent Services Ecosystem* project was funded under the EU H2020 TRINITY Project Open Call 2 (GA nr 825196), with the aim of developing a set of applications, specified by world leading robot manufacturers, integrating them into the MYWAI EaaS Platform for edge, fog or cloud delivery, but also into existing PLC and SCADA systems.

The objective of the RAISE™ project (***R**obots **A**s an **I**ntelligent **S**ervices **E**cosystem*) is to enable the *servitization* model for robotic manufacturers, which, although experiencing good sales performances in large industries, could be interested in offering an innovative *Robots-as-a-Service (RaaS)* model, to target SME "long tail" market companies, subscribing on an outcome-based model, see Fig. 6.

Fig. 6. The RAISE™ RaaS project

This RaaS model will enable businesses to outsource a capital-intensive part of their assets and offload risks: e.g. zero upfront expenses. In addition, it will also enable organizations monitoring their robotic workforces, to readily and rapidly scale up and down in response to customer demands and market conditions. Finally, it will enable new, optimized models of insurance, e.g. parametric plans, extended warranties, etc.

A demonstrator of the proposed RaaS solution for Mitsubishi Electrics™, one of the world leading robot manufacturers was developed, with the engagement of Italian and Lithuanian start-ups, YOLO™ S.R.L. and UAB NeuroTechnologijos™, respectively, for the independent provisioning of Fin-tech, Insur-tech and Service-tech services. The integration and interoperability of different providers for selected services was achieved through the adoption of the new machinery companion specifications made available by the OPC UA foundation developing Industry 4 interoperability standards.

The proposed solution has been designed to reduce SME's effort to embrace robotic workforces by 50% at regime whilst introducing new affordable and effective business models, insuring machinery on produced outcomes and paying its cost into recurring OpEx services.

In summary, the resulting RAISE™ RaaS platform will be a world premier thanks to its blend of Edge AI algorithms, after market maintenance and warranty extension services, as well as innovative Fin-tech and Insur-tech service models.

As a further opportunity for investigating and enforcing the *servitization* concept, we have also been involved in the *PromenAIde*™ - ***PRO****gnostic* ***M****edical* ***E****quipment maintenance services* ***EN****abled by* ***AI*** *on* ***D****LT and* ***E****dge* project, funded by the Italian Liguria Region under the POR-FESR 2014-2020 program.

This project targets the design and development of a platform for the *servitization* of products in the MedTech sector, where the recent pandemic has accelerated the need to service-oriented models (see [10] for a framework for assessment and implementation of product-service systems strategies in health-care), capable of improving equipment performance, which in this case are ultrasound and MRI scanners, provided by the world-leading ESAOTE™, see Fig. 7.

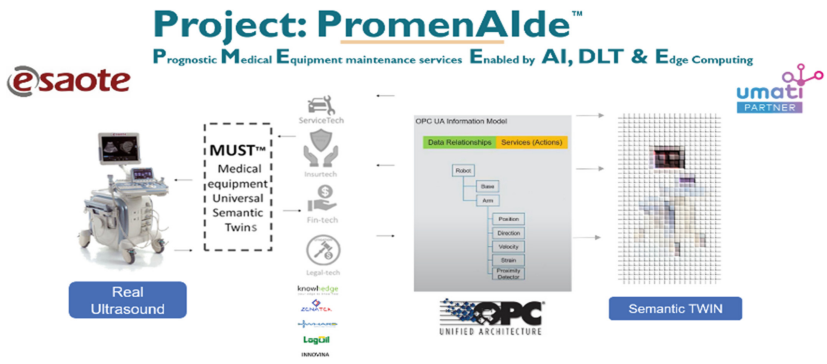


Fig. 7. The PromenAIde project

The project aims at establishing an efficient and secure ecosystem, able to continuously offer value-added services, integrated with medical equipment already available on the market, thanks to recent technological advancements, designed for certification and payment distributed in cryptocurrency even directly between intelligent machines (M2MP, or "Machine-to-Machine Payments").

The realization of the PromenAIde™ solution will rely on a micro-services architectural design and on the adoption of the API-FIRST paradigm, which implies prioritizing the formalization of interfaces to third-party systems and the standardization of data/information models at the semantic level, by involving emerging specifications like OPC UA, to promote the development of additional interfaces for customers willing to exploit them in different industrial sectors and for the creation of new third parties added-value services.

4 Conclusion

In this paper, the concept of CPE-aaS is explored, together with the role of IoT edge platforms in their implementation. To illustrate our approach, two successful projects were briefly introduced: RAISE and PromenAIde. The focus of our approach is that OEMs can benefit from the proposed edge platform, to avoid starting from the scratch setting up their servitization platform. Instead, they can use the proposed platform which provides all required features to get started, being able to scale up and down easily in a later stage. Servitization requires a decentralized approach, to collect all the necessary data to provide the best possible service. In this regard, we can see that cyber-physical technologies are a requirement of servitization, as a crucial enabler, and not an end.

References

1. Wiener, N.: Cybernetics: or Control and Communication in the Animal and the Machine, 2nd edn. MIT Press, Cambridge (1948)
2. Piano Nazionale di Ripresa e Resilienza, Mise. https://www.mise.gov.it/index.php/it/68-incentivi/2042324-piano-nazionale-di-ripresa-e-resilienza-i-progetti-del-mise. Accessed 7 June 2022
3. Mastrogiacomo, L., Barravecchia, F., Franceschini, F.: A worldwide survey on manufacturing servitization. Int. J. Adv. Manuf. Technol. **103**(9–12), 3927–3942 (2019). https://doi.org/10.1007/s00170-019-03740-z
4. Smith, D.J.: Power-by-the-hour: the role of technology in reshaping business strategy at Rolls-Royce. Technol. Anal. Strateg. Manag. **25**(8), 987–1007 (2013). https://doi.org/10.1080/09537325.2013.823147
5. Bianchi, N.P., Evans, S., Revetria, R., Tonelli, F.: Influencing factors of successful transitions towards product-service systems: a simulation approach. Int. J. Math. Comput. Simul. **3**(1), 30–43 (2009)
6. Lah, T., Wood, J.B.: Technology-as-a-Service Playbook: How to Grow a Profitable Subscription Business. Point B, Inc., Seattle (2016)
7. Creating New Business Models with Design Thinking. Equipment Finance Advisor. https://www.equipmentfa.com/articles/31948/creating-new-business-models-with-design-thinking. Accessed 2 June 2022

8. Tonelli, F., Evans, S., Taticchi, P.: Industrial sustainability: challenges, perspectives, actions. Int. J. Bus. Innov. Res. **7**(2), 143–163 (2013)
9. Baines, T., Lightfood, H.: Made to Serve: How Manufacturers Can Compete Through Servitization and Product—Service Systems. Wiley (2012). https://doi.org/10.1002/9781119207955
10. Tonelli, F., Taticchi, P., Starnini, S.E.: A framework for assessment and implementation of product-service systems strategies: learning from an action research in the health-care sector. WSEAS Trans. Bus. Econ. **6**(7), 303–319 (2009)
11. Ardolino, M., Rapaccini, M., Saccani, N., Gaiardelli, P., Crespi, G., Ruggeri, C.: The role of digital technologies for the service transformation of industrial companies. Int. J. Prod. Res. **56**(6), 2116–2132 (2018). https://doi.org/10.1080/00207543.2017.1324224
12. Dinges, V., Urmetzer, F., Martinez, V., et al.: The future of servitization: technologies that will make a difference. Cambridge Service Alliance, University of Cambridge, Cambridge (2015)
13. Kuhlenkötter, B., et al.: New perspectives for generating smart PSS solutions – life cycle, methodologies and transformation. Procedia CIRP **64**, 217–222 (2017). https://doi.org/10.1016/j.procir.2017.03.036. ISSN 2212-8271

Structural Health Monitoring

A Sensing System for Defects Detection in Reinforced Aluminium Panels by Ultrasonic Guided Waves Sensors

Lorenzo Capineri(✉), Andrea Bulletti, Cosimo Damiani, and Luca Bergamaschi

Università degli Studi di Firenze, 50139 Florence, Italy
lorenzo.capineri@unifi.it

Abstract. The new living modules of spacecrafts are made of reinforced aluminium panels of large dimensions and are eventually subjected to impacts with space debris causing damages. The defect detection presented in this paper is based on the design of piezopolymer interdigital transducers (IDT) based on an analytical model that predicts the optimal frequency and placement on the reinforced panel to avoid the influence of the reinforcing ribs on the ultrasonic guided wave propagation. The paper presents the experimental validation of the method on a representative panel with regular ribs distance of 10 cm and with IDT operating at 650 kHz. The layout of the sensing system is proposed to be integrated in a multichannel electronic system to demonstrate the feasibility for the applications in Structural Health Monitoring (SHM). Finally tests on standard surface defects are reported.

Keywords: Aerospace · Defect · Aluminium reinforced panel · Ultrasonic guided waves · Interdigital transducers · Beat-length

1 Introduction

The increasing duration of space missions characterized by the prolonged presence of astronauts, requires further steps towards an on-board increase of safety both for orbiting modules and future surface habitats in permanent outposts. It is necessary to be able to locate and quickly intervene in case of MMOD (Micro-Meteoroids & Orbital Debris) impacts or accidental damaging, with repairing the habitable pressurized volumes of the metallic structures with integrated structural reinforcement elements that are the current baseline design solution. Furthermore, with the need to increase the size of manned modules to provide more comfort on-board, the monitoring aspect is going to become a must to efficiently allow the rapid crew action on damages.

For the Cis-Lunar Space Station modules, in the frame of the Lunar Gateway program, the need for a SHM is specifically addressed by NASA in a dedicated document GP 10055 - Gateway Program Specification for Structural Health Monitoring, to underline that SHM systems will be required.

A preliminary investigation has been performed by assessing the Ultrasonic guided waves (UGW) propagation through a reinforced aluminum panel with the implementation on a dedicated test object.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 331–339, 2023.
https://doi.org/10.1007/978-3-031-16281-7_32

2 The Aluminium Reinforced Panel Adopted

Ultrasonic guided waves (UGW) are one of the main methods adopted for defect detection in nondestructive testing (NDT) and SHM [1]. Their capability of operating at low frequencies (100 kHz–1 MHz) with wavelengths comparable to defect size and position across the laminate thickness has granted several applications in aerospace, defense, energy transportation and conversion system with relatively low-cost electronic instruments [2]. Moreover, the low attenuation of the UGWs provides efficient system solutions with a low density of transducers with advantages on cabling, wireless connections, and power consumption [3]. In this paper we tackle the problem of defect detection in aluminum panels designed to fabricate living modules for new spacecraft employed in future mission. While the lightness requirement is obtained with aluminum metal with low thickness (order of mm), the mechanical stress is achieved with reinforcing ribs. Such geometry differs substantially from the standard conditions for UGW that are planar laminates. An example of such panels is described in Fig. 1.

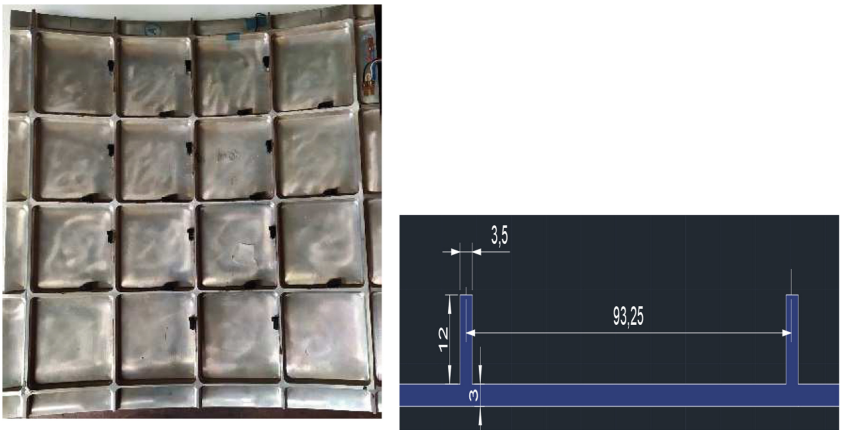


Fig. 1. (Left) Bottom view of the reinforced aluminum panel with regular distance between ribs of 100 mm. Area 420 mm × 420 mm. (Right) Section view of the panel planar surface with dimension in mm. Material Al T6; thickness t = 3 mm; distance between ribs d = 93,25 m; ribs' dimension 12 mm × 3.5 mm.

The propagation of ultrasonic wave is dispersive and for the planar structure the phase velocity depends on the frequency - thickness product. Figure 2 shows the simulated phase velocities curves for an aluminum panel type 6060 alloy.

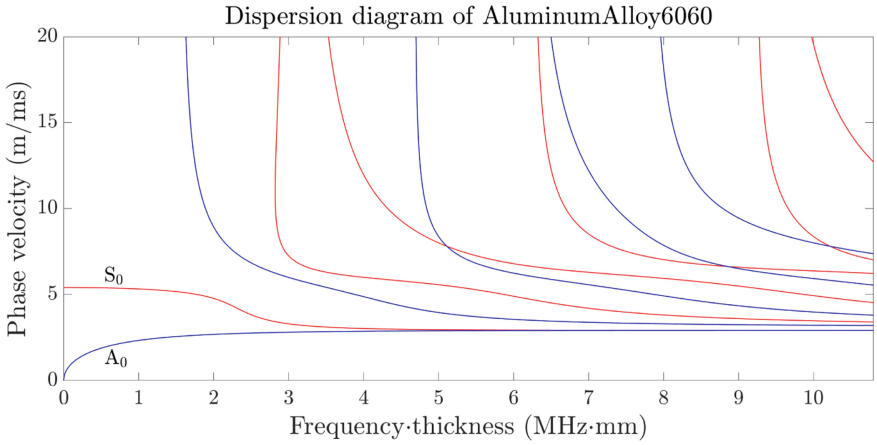


Fig. 2. Dispersion curves for the planar aluminum panel obtained by the Disperse Calculator [4].

3 The Analytical Model Based on the Beat-Length for the IDT Design

The main problem encountered for the NDT inspection of a reinforced panel is the evaluation of the influence on the propagation of the UGW through the ribs, as they represent a significant variation of the thickness t of the planar structure. According to recent studies published by B. Massery and P. Fromme [5] and by Hlavaty et al. [6], is it possible to combine different type of guided modes (A_0 and S_0) to generate quasi-Rayleigh waves which have a minimum in the out of plane displacement at the position of the ribs. Consequently, the influence of the ribs on the UGW is minimized while the increased amplitude in the squared area enclosed by the ribs is high. The combination of two modes propagating with their characteristics phase velocity creates a beating effect and the composite wave have a spatially period called "Beat-Length", that is defined by the following relationship:

$$L_B = \frac{2\pi}{k_{A_0} - k_{S_0}} = \frac{\lambda_{S_0} * \lambda_{A_0}}{\lambda_{S_0} - \lambda_{A_0}} \tag{1}$$

where λ_{A0} and λ_{S0} are the wavelength of the two modes A_0 and S_0 respectively at a defined frequency f_0.

According to (1) the distance d of the reinforcing ribs must be an integer multiple n of the beat-length LB:

$$d = n * L_B \tag{2}$$

From this analytical model we can derive information necessary to design the transducer operating at a certain frequency and with assigned distance from the ribs. The simulations reported in [7] show that for increasing of n, the out of plane amplitude on the ribs' positions tend to increase and the mitigation effect decreases. For larger n value the area coverage is achieved with a lower transducer density.

In Table 1 are reported possible choices for the operating frequency f_0 of the UGW transducers calculated with phase velocities reported in Fig. 2:

Table 1. Selection table for operating frequency.

f_0 [MHz]	L_B [mm]	n
1.735	*90*	*1*
1.385	*45*	*2*
1.2	*30*	*3*
1	*18*	*5*
0.55–0.75	*9*	*10*

According to Table 1, we decided to operate the transducers in the bandwidth 0.55–0.75 MHz corresponding to a distance $d = n*LB = 90$ mm almost equal to the distance between ribs. The lower frequency corresponds also to wavelengths comparable to defect size which improves the detection capability; however higher frequency (e.g. 1 MHz) could be experimented as they improve the resolution for the defect detection.

4 IDT Design and Fabrication

The theory reported in the previous section implies the realization of UGW transducers capable to select the desired modes at the specified frequency. The IDT type comes in help for these requirements because the interdigital electrodes geometry can be optimized for this aim [8, 9]. The design of the transducer starts with the geometry definition (eg. AutoCAD model) to be transferred for any possible fabrication technology. The modes selection can be obtained by the phase velocity diagram and considering the following relationship between wavelength, frequency, finger spacing:

$$V_p/(f_0 * t) = l_0/t \tag{3}$$

where V_p is the phase velocity and t the laminate thickness.

This relationship can be reported in the dispersion curves plots being a straight line with slope depending on the selected wavelength λ_0. The interdigital finger spacing S is half of λ_0. Our design started with the selection of $\lambda_0 = 4$ mm for the A_0 mode, being the propagating mode with larger amplitude and capable of great sensitivity of surface disturbance due to a defect. Figure 3 shows the electrode geometry and in Fig. 4 the corresponding modes excited in the frequency bandwidth 550–750 kHz.

Finally, the fabrication was carried out using a 110 μm thick PVDF film with sputtered gold metallization. The AutoCAD model was transferred to a Nd-Yag laser equipment for the gold metal ablation. The piezopolymer film is the preferred technology as it is capable to sustain large vibration of the investigated structure without damage and shows a good sensitivity in the receiving mode. For the piezopolymer film mechanical coupling with the panel, a bi-adhesive tape is adopted that allows also easy adjustment

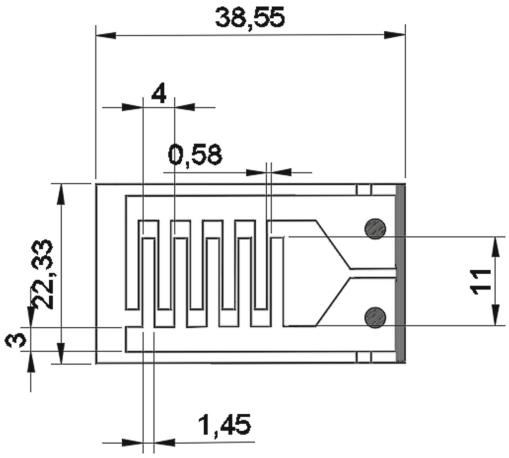


Fig. 3. Electrodes geometry for the IDT. Dimensions in [mm].

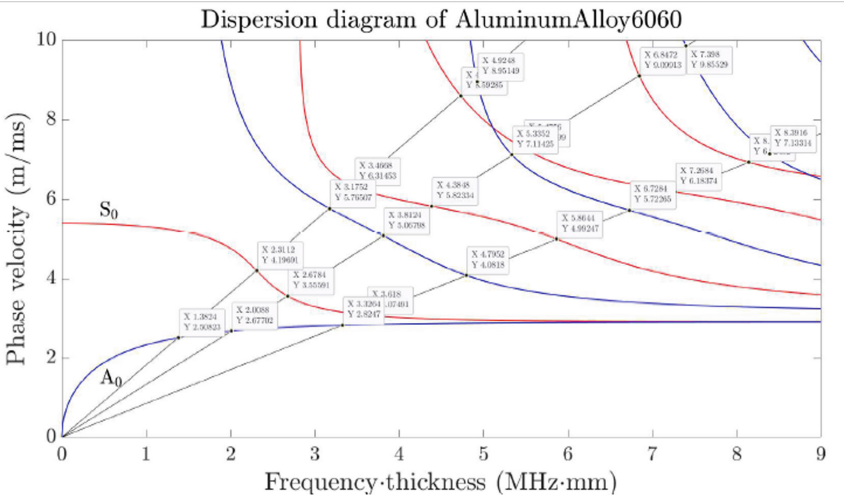


Fig. 4. Selection of modes with designed IDT in the frequency range 550–750 kHz for a 3 mm thick aluminum laminate.

of the sensor position without spoiling the surface as in the case of permanent bonding with epoxy (Fig. 5).

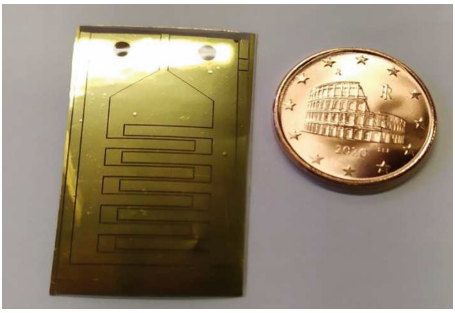


Fig. 5. IDT fabricated with a piezopolymer (PVDF) film.

5 Validation of the IDT Response in Pitch and Catch Configuration on the Reinforced Aluminium Panel

A pair of IDTs were installed at the defined beating distance of 9 mm from the ribs and the distance d_2 between the IDT equal to 206 mm as shown in Fig. 6.

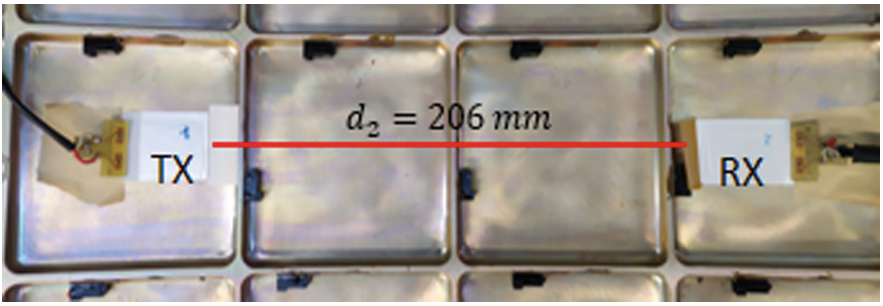


Fig. 6. IDT placement on the reinforced aluminum panel at distance defined by the beat-length criterion.

The transmitting IDT (TX) is excited with a linear power amplifier with a 4 cycle -sinewave burst at 650 kHz with amplitude 90 Vpp. The receiving element is connected to a differential low noise voltage mode amplifier with gain 60 dB in the bandwidth 100 kHz–1 MHz. The received signal was digitized and displayed for different excitation frequency to validate the optimal operating frequency, the bandwidth, and the validity of the assumption of the design based on the beat-length. The picture in Fig. 7 shows a main signal corresponding to the A_0 mode at 650 kHz.

Different frequencies down to 350 kHz and up to 800 kHz were explored and it is confirmed the very high attenuation due to the reinforcing ribs. From the point of view of the sensing system installed on a large panel with surface up to 12 square meter is essential to evaluate the attenuation for the different frequency in the IDT bandwidth. By comparing the signal at 650 kHz acquired at a distance about half of the $d_2 = 206$ mm (see Fig. 6), we found an attenuation of about 6 dB. Then for covering about 1 m, we expect with a narrow beam IDT to get an overall attenuation of about 30 dB.

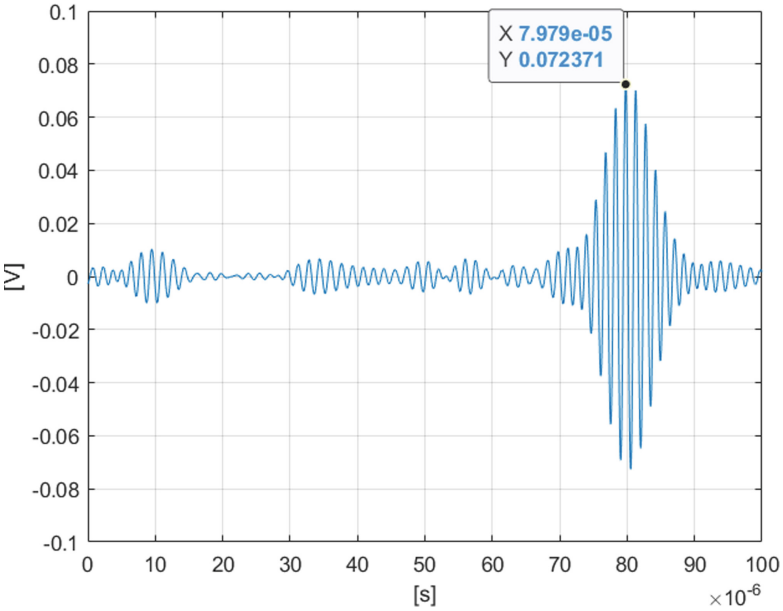


Fig. 7. Received signal travelling across two reinforcing ribs after band pass filtering (central frequency 600 kHz and bandwidth 100 kHz). The main signal at about 80 μs with about 0.14 Vpp corresponds to the A_0 mode. The faster signals at about 35 μs and 55 μs correspond to S_0 and Shear Horizontal (Rayleigh) modes respectively.

6 Validation of the IDT in Pitch and Catch Configuration for Defect Detection

The final goal of this study is the detection of representative defects for aerospace structure monitoring. Among several defects we selected one type that can be generated during the service of a structure and the shape is a semi-spherical shape with depth 1.65 mm and same width = 1.65 mm. The realization of the defect placed on the axis of the two transducers is shown in Fig. 8.

By comparing the signal with a baseline acquired on a pristine panel is possible to apply simple detection algorithm based on amplitude threshold or on more sophisticated algorithm like the composite correlation index proposed by the authors [10]. The next picture shows both signals were a different of more than 16% in the peak-to-peak amplitude can be observed. It is important to outline that the defect dimensions are challenging as they are a fraction of the $\lambda_{A0} = 4$ mm (Fig. 9).

Fig. 8. Artificial defect placed on the axis of the IDT pair is marked with a red circle.

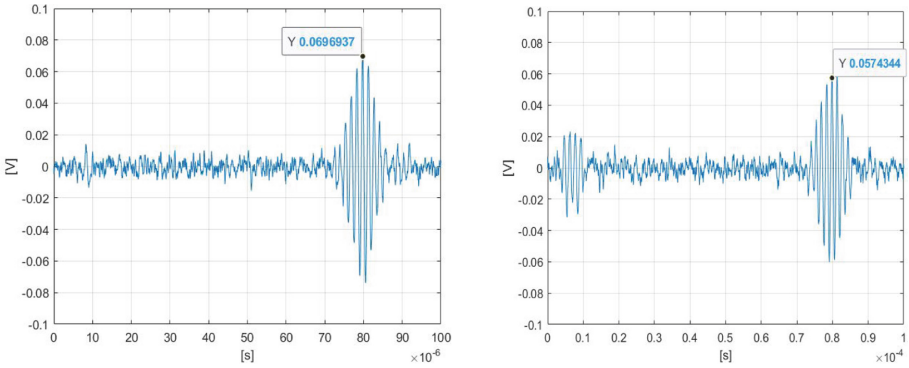


Fig. 9. (Left) A_0 mode at 650 kHz on a pristine panel and (Right) when A_0 mode is propagating through the artificial defect. The start signal corresponds to the electromagnetic coupling between and electric system.

7 Conclusions

The work presented in this paper proposes a methodology for designing intelligent SHM systems for active damage monitoring of space modules made of reinforced aluminium panels. The novelty of the work is the design and fabrication of piezopolymer IDTs that can select the required Lamb wave modes in order to propagate through the reinforcing ribs with low attenuation of the ultrasonic signal. By proper placing the IDT respect to the ribs the defect detection capability has been demonstrated operating at 660 kHz. Further works will be the investigation of higher frequencies to improve the resolution of defect detection without compromising the signal to noise ratio.

References

1. Raghavan, A., Cesnik, C.E.S.: Review of guided-wave structural health monitoring. Shock Vib. Dig. **39**, 91–114 (2007)

2. Rose, J.L.: Ultrasonic Guided Waves in Solid Media. Cambridge University Press, New York (2014). https://doi.org/10.1017/CBO9781107273610
3. Kino, G.S.: Acoustic Waves: Devices, Imaging, and Analog Signal Processing. Prentice-Hall, Englewood Cliffs (1987)
4. Huber, A.: Dispersion Calculator (2018)
5. Masserey, B., Fromme, P.: Surface defect detection in stiffened plate structures using Rayleigh-like waves. NDT E Int. **42**(6), 564–572 (2009). https://doi.org/10.1016/j.ndteint.2009.04.006
6. Hlavatý, M., Starek, L., Musil, M., Hučko, B.: Ultrasonic defect detection of structural plates using quasi-Rayleigh waves. Strojnícky Časopis - J. Mech. Eng. **67**(2), 37–50 (2017). https://doi.org/10.1515/scjme-2017-0016
7. Masserey, B., Fromme, P.: On the reflection of coupled Rayleigh-like waves at surface defects in plates. J. Acoust. Soc. Am. **123**(1), 88–98 (2008). https://doi.org/10.1121/1.2805668
8. Capineri, L., Gallai, A., Masotti, L.: Design criteria and manufacturing technology of piezopolymer transducer arrays for acoustic guided waves detection. In: 2002 IEEE Ultrasonics Symposium, 2002, Proceedings, Munich, Germany, pp. 857–860 (2002). https://doi.org/10.1109/ULTSYM.2002.1193532
9. Bellan, F., et al.: A new design and manufacturing process for embedded Lamb waves interdigital transducers based on piezopolymer film. Sens. Actuators Phys. **123–124**, 379–387 (2005). https://doi.org/10.1016/j.sna.2005.05.013
10. Poddar, B., Giurgiutiu, V.: Complex modes expansion with vector projection using power flow to simulate Lamb waves scattering from horizontal cracks and disbonds. J. Acoust. Soc. Am. **140**(3), 2123 (2016). https://doi.org/10.1121/1.4963087

A Tiny Machine Learning Approach to the Edge Localization of Acoustic Sources via Convolutional Neural Networks

Federica Zonzini[1(✉)], Giacomo Donati[2], and Luca De Marchi[2]

[1] Adavanced Research Center on Electronic Systems "Ercole De Castro" (ARCES), University of Bologna, 40136 Bologna, Italy
federica.zonzini@unibo.it

[2] Department of Electrical, Electronic, and Information Engineering (DEI), University of Bologna, 40136 Bologna, Italy
{giacomo.donati5,l.demarchi}@unibo.it

Abstract. Source localization is a critical step in Acoustic Emission (AE)–based Structural Health Monitoring (SHM), since it allows to identify the point of a structure where most of the acoustic activity is growing due to both ageing (e.g., cracks, delamination, etc.) and sudden flaws. Recently, Artificial Intelligence (AI) algorithms have been proposed, which can overcome standard statistical methods especially when the signal–to–noise ratio is poor. In this work, the embodiment of tiny Convolutional Neural Network (CNN) models on a 32–bit microcontroller unit is presented for the task of Time of Arrival (ToA) estimation, which is the crucial parameter to be estimated for AE localization. Experimental results on real–field data prove that the embedded models can achieve satisfying accuracy for AE identification.

Keywords: Acoustic source localization · Convolutional neural network · Structural health monitoring · Tiny machine learning

1 Introduction

Acoustic Emission (AE)–based monitoring represents one of the most effective Non Destructive Evaluation (NDE) approaches for the Structural Health Monitoring (SHM) of structures or materials subject to stress [7,8]. The underpinning principle behind AE is that the occurrence of acoustic events such as cracks, delaminations, disbonding, etc. is intrinsically related to the status of integrity such that the higher the frequency and the intensity of the recorded acoustic activity, the higher the level of potential structural degrade. Beside the mere AE activity tracking, one of the most important advantages of AE over alternative monitoring methods relies on the possibility to localize such sources by passively capturing the acoustic response of the structure. A crucial step in this direction is the proper estimation of the Time of Arrival (ToA), namely the time at which

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 340–349, 2023.
https://doi.org/10.1007/978-3-031-16281-7_33

the induced mechanical wave released at the occurrence of the acoustic event arrives to the acquisition unit [6]: indeed, this parameter is required by most of the established AE signal processing techniques.

The taxonomy of the strategies proposed for the task of ToA estimation is very broad and spans from statistical methods to artificial intelligence (AI) solutions [11], the latter being an emerging trend of research in recent years thanks to their superior ability in learning very complicated patterns hidden within data. Indeed, compared with conventional techniques, among which the Akaike Information Criterion can be listed, machine learning methods are superior in that they can be applicable even when the Signal–to–Noise Ratio (SNR) conditions are very unfavourable, either as a consequence of electromagnetic noise and rubbing disturbances in the surrounding of the monitored environment or due to the noise affecting the employed electronics [1].

However, to attain sufficient estimation accuracy, most of these AI methodologies are very onerous in terms of computational power and model size and, thus, they are typically deployed in remote servers and executed in a time and power–hungry manner. Conversely, in order these Neural Network (NN) models to be applicable in practical long–term monitoring scenarios, they should be run on edge sensors so that it is possible minimize the payload to be transmitted over the monitoring network. In this scenario, in fact, just few scalars (the ones related to the ToA) rather than the entire raw waveforms are transmitted, and this, in turn, reduces the overall system latency. Noteworthy, this requirement of network load reduction is crucial in wireless monitoring systems, where significant constraints must be satisfied in terms of power consumption and channel bandwidth in order to avoid network congestion and ensure long–lasting serviceability.

Therefore, a pivotal challenge to be faced for the development of the next generation of edge–oriented AE–driven SHM architectures, is the embodiment of AI models in resource–constrained devices [4]. Enabling edge inference represents a radical shift of perspective, that is inspired by the novel and pioneering concept of Tiny Machine Learning (TinyML), which is defined as the capability of running AI at the boundary between the physical and the digital world[1], i.e., by means of edge or extreme edge sensors, in a power and computationally efficient manner. Note that this TinyML approach is remarkably different from most of the currently adopted AE monitoring strategies based on commercial solutions, which still depend on very bulky, costly and monolithic instrumentation, that prevent their applicability for real–time on–condition maintenance.

This works aims at bridging this gap by demonstrating the actual embodiment of tiny CNN models for ToA estimation on a resource–constrained board equipped with a low–power and low–cost microcontroller unit, which is actually in charge of running the NN model in a self–contained manner. The results of the prediction are used in the framework of an experimental setting for AE source localization.

The content of the work is organized as follows. In Sect. 2, the CNN models and the different quantization strategies necessary for their porting are

[1] https://www.tinyml.org/.

presented. Then, the materials and methods employed for the sake of model training and validation are described in Sect. 3, while experimental results are discussed in Sect. 4. Conclusions end the paper.

2 Tiny CNN Models for ToA Estimation @ Edge

Two different Convolutional Neural Network (CNN) models for ToA estimation are considered for porting on embedded devices. *Model 1* was presented in [11], and consists of four successive 1-Dimensional convolutional (CNN1D) layers plus stacked max pooling layers, followed by a global average pooling layer and two fully connected layers (see Fig. 1(a)) for a total amount of 134 481 parameters. *Model 2* has been designed to be a tinier version of the previous one and presents the structure drawn in Fig. 1(b): as can be seen, the major differences pertain to the reduction in the number of filters per layer and in the dimension of the kernel size, that has been decreased to 8 (apart from the first CNN1D layer which is identical between the two models). The total number of parameters of this second model equals to 76 715.

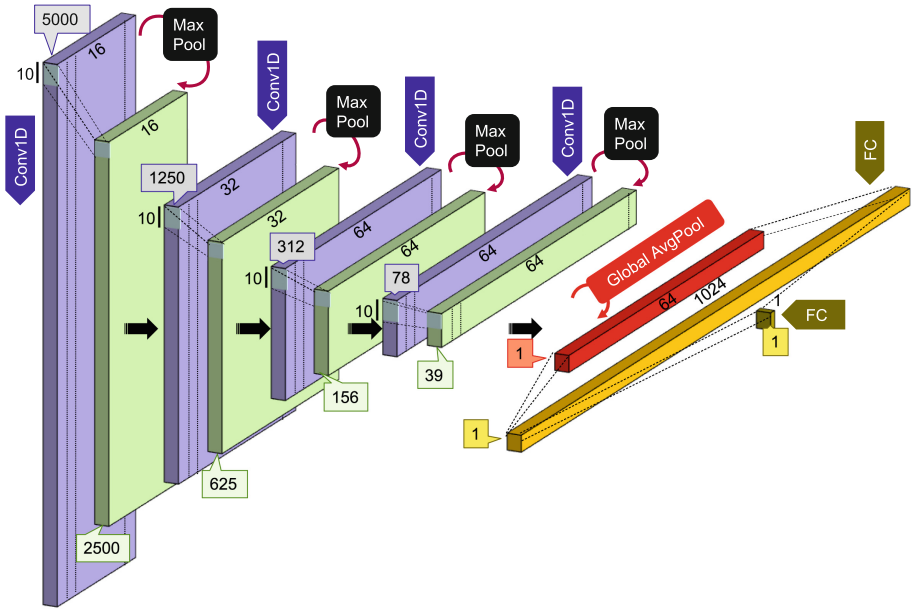
These models were firstly trained offline using the Keras Python API in the TensorFlow platform[2]. Then, following the standard TinyML deployment process recommended in [10], a quantized version has been obtained and uploaded to the target board, where the inference has finally been executed.

2.1 Quantization Techniques

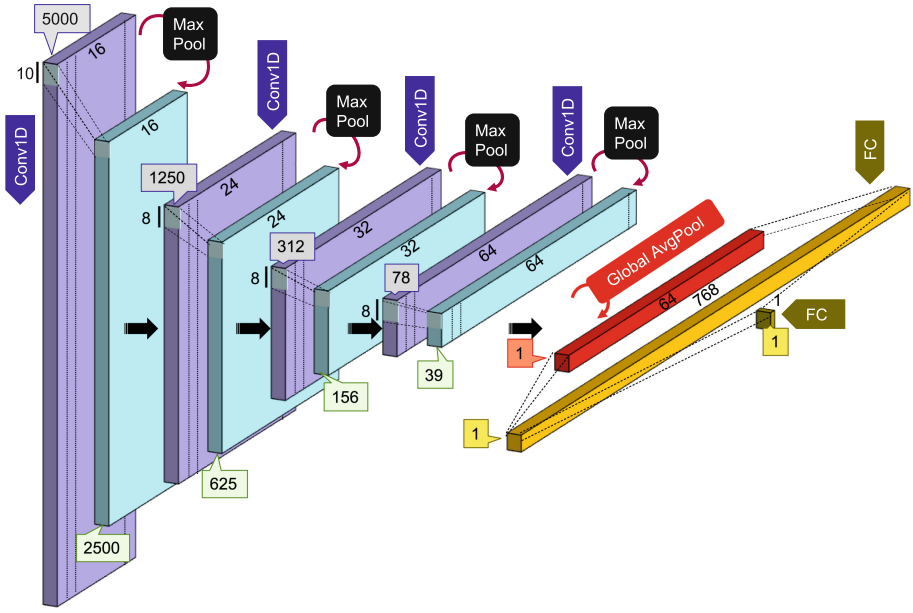
Different quantization techniques have been adopted to fit the above–described NN models into the limited computing and memory capabilities of low–power processors. The adopted TF platform supports two main quantization techniques: *post–training* quantization and *quantization–aware* (Q–aware) training. In both cases, once the model has been designed, a conversion to the TensorFlow Lite (TFlite) format has to be performed in order to obtain a quantized version of it, suitable for deployment on a generic edge–device runtime. In particular, we focused on full integer conversion strategies, hence comparing their related performances to those relative to high–depth bit quantization models, i.e., to the original ones executed in Keras with 32–bit representation and floating point input/output data, to which we will refer to as `Float32` in the next sections. We can summarize such explored methodologies as follows:

- *Post–training full integer quantization* (`Int8`): all the mathematical functions to be realized are implemented via full integer quantized operators. Accordingly, weights and activation have 8–bit precision, including input and output tensors.
- *Post–training integer quantization with float fallback* (`Int8 + Float` fall–back): the rationale behind this procedure is analogous to the `Int8` solutions, the main differences being in the fact that i) float operators are used when an

[2] https://www.tensorflow.org/.



(a) Model 1



(b) Model 2

Fig. 1. CNN models for ToA estimation: (a) original model (Model 1) and (b) its shrunk version (Model 2).

integer implementation for the same function is not available and ii) higher precision bit–depth (i.e., 32–bit) is admitted for input/output tensors. Given these characteristics, such a quantization strategies creates a hybrid model between the full `Int8` and full `Float32` solutions.
– *Pre–training quantization* (Q-aware): the training process is *quantization–aware*, in the sense that it emulates inference with 8–bit quantization using precision reduction in weights and activation representation, such that the performance degradation after deployment can be prevented. Despite the above mentioned benefit, this quantization method could lead to a slower convergence time during training. Moreover, similarly to the post–training quantization discussed above, input and output tensors could be formatted as integer 8–bit values (Q–aware `Int8`) or converted to floats (Q–aware + Float I/O).

3 Model Deployment Process, Training and Testing

3.1 Materials

The STM32F407G DISC-1 [9] board equipped with an ARM-Cortex M4 capable of running with a clock frequency up to 168 MHz has been employed for prototyping purposes. It features 1 MB of FLASH memory and 192 KB of SRAM, which are compatible with the typical characteristics of edge nodes to be deployed for long–lasting SHM monitoring. Moreover, the X–CUBE–AI expansion package for the STM32XCube has been used: this AI ecosystem has been exploited for the embodiment of the models into the selected platform.

3.2 Dataset Generation

The dataset for the training phase has been built via numerical simulations. This procedure based on synthetic AE signals has been preferred over a purely experimental approach for two main reasons: the former is that it allows to speed up the data collection phase, since a relatively short amount of time is necessary to generate a representative pool of data; secondly, and more importantly, it permits the creation of very accurate ground truth labels for supervised learning (as it is the case of CNN–based solutions), a condition which is not applicable in passive AE monitoring scenarios where the true AE triggering time is always unknown to the sensing system, which is only programmed to trigger acquisition upon crossing of a pre–selected energy threshold in the incoming wave. This is due to the fact that the detected waves are generated by the structure itself due to cracks growth, undesired impacts, or corrosion. Hence, they are not controllable and not predictable in advance and can only be characterized in post–processing. Therefore, for the prospective application scenarios dominated by the absence of a proper synchronization between reception and actuation, it is not possible to derive the true ToA label in an experimental manner.

When an AE event occurs as a consequence of crack, corrosion or delamination processes in a waveguide, ultrasonic guided waves (GWs) are generated

and can travel long distances. The propagation characteristics of GWs can be numerically modeled when the geometrical and mechanical parameters of the monitored structure are known. This is indeed one of the utmost advantage in pursuing this numerical simulation strategy. To this end, an *ad–hoc* ray–tracing algorithm has been implemented, which simulates the peculiar dispersive propagation behavior of GWs between different points of a thick aluminum plate while taking into consideration also the effects of multiple reflections due to the mechanical boundaries. As already described in [11], the entire number of combinations between the selected points for AE actuation and reception has been simulated, while also changing the SNR (uniformly increased from 2 dB to 30 dB at integer steps of 1 dB) for a total amount of 60 000 time series: 80% of this data was used for training and validation and the remaining 20% was allocated to testing.

3.3 Validation Process

The performance of the differently quantized models has firstly been assessed on the synthetic testing dataset (12 000 time series) introduced in Sect. 3.2. This initial validation process has been necessary to evaluate the effect of the quantization itself on the prediction capability of the CNNs when running on the edge sensor.

Then, the ToA prediction algorithms have also been tested for localization purposes in an experimental setting involving a thin aluminum plate ($1000 \times 1000 \times 3$mm) instrumented with three custom sensor nodes (installed on three corners of the structure) developed within the Intelligent Sensor Systems lab of the University of Bologna and located at as many corners of the structure (see Fig. 2). Thanks to the compact design including all the circuitry and electronics necessary to collect, pre-process and characterize signals, each device works as a miniaturized oscilloscope capable of acquiring, at the same time, signals on three different input channels with a capacity of 4MS/s. All the details about the sensor node characteristics can be found in [3,12]. This plate has been selected since it shares the identical mechanical properties with the one adopted for numerical simulations and, thus, it allows the same NN models to be trained on the simulated time series to be exploited for the purpose.

Nine different positions have been considered for excitation: each test has been repeated three times, for a total amount of 27 tests. The adopted sensor installation plan is compatible with the triangulation method in [5], whose complete mathematical formulation can be found in [11]: the algorithm is advantageous in that, thanks to simple geometrical considerations, it allows the retrieval of a source position simply by knowing the ToA of three sensing units placed at known position. Such testing procedure as been necessary since, as anticipated, it is not possible to get an educated guess about the true label in case of real–field scenario; therefore, the rationale behind the performed experiments is that achieving good source localization accuracy can be considered as an indirect means for ToA prediction validation.

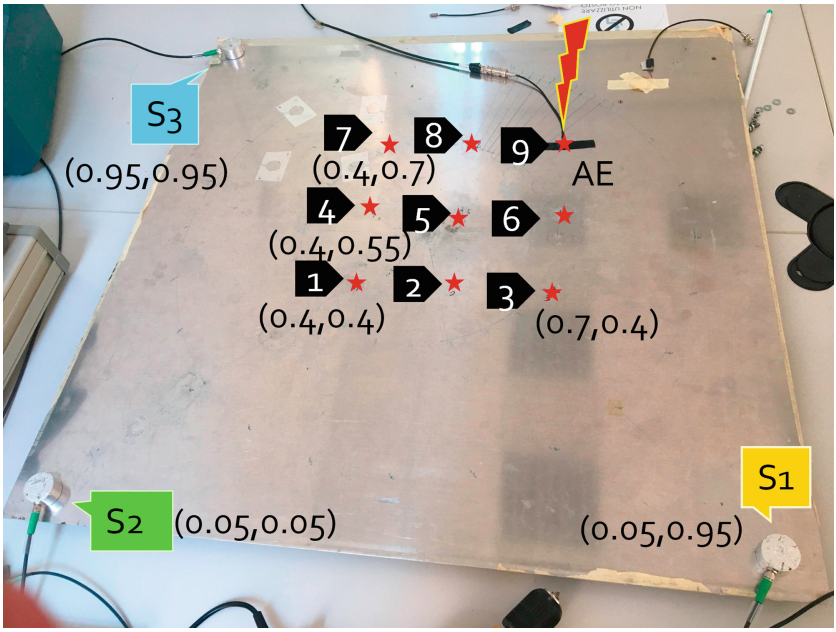


Fig. 2. Experimental setup employed for AE source localization: three sensors (S1, S2, S3) are installed on three corners of the plate, while 9 different points equally spaced are considered for AE actuation.

Furthermore, as anticipated, since the primary advantage of AI approaches is that they can efficiently handle even very noisy AE monitoring scenarios, the impact of progressively increasing noise levels on the edge inference has been specifically evaluated. To this end, gathered data have been corrupted with an additive white Gaussian noise such that the corresponding SNR moves from 20 dB to 4 dB at integer steps of 4 dB. Despite the fact that the nature of the background noise of real AE signals can indeed differ [2], additive white stationary noise (such as the one generated by electronic components) can be assumed to be the main source of SNR degradation and, consequently, was used to simulate noisy AE scenarios in this study.

4 Results

4.1 Preliminary Validation on Synthetic Signals

Results after testing the CNN models and corresponding quantization strategies introduced in Sect. 2 on synthetic data are summarized in Table 1. Here, the (percentage) degradation in the Mean Absolute Error (MAE) (computed as the magnitude of the difference between the predicted ToA and the true ToA label) with respect to the one related to the original unquantized model running in

Keras (Model 1), indicated as $\text{MAE}_{\text{Model1Keras}}$ and amounting to 12.39 µs, is reported and computed as

$$\text{MAE loss}\% = 1 - \frac{\text{MAE}_{\text{q}}}{\text{MAE}_{\text{Model 1 Keras}}} \cdot 100 \tag{1}$$

where MAE_{q} is the MAE (in µs) for the generic quantized versions of the models described in Sect. 2.1. Additionally, the computational complexity, intended as the number of Multiply–and–Accumulate (MAC) operations, has also been estimated, together with the overall execution time and the memory occupancy.

As can be observed by analyzing the performances related to Model 1, the advantage of Q–aware solutions is twofold: i) they require much less memory space as granted by the low–depth quantization and ii) their accuracy degradation is one order of magnitude lower to the one obtained with post–training quantization. Leveraging this initial outcome, the Q–aware quantization is the only method adopted for Model 2, which shows a very minor loss in the overall prediction accuracy, with the crucial benefit of halving both the execution time, which moves from averagely 410 ms to nearly 240 ms, and the memory occupancy (77 KB vs 134 KB).

For these reasons, only Model 2 and its Q–aware variants will be taken into consideration in the following analyses.

Table 1. Performance metrics on the synthetic dataset for the implemented CNN model and relative quantization strategies while running on the L4 MCU.

Model/Quantization Type	Complexity [MAC]	SRAM [KB]	FLASH [KB]	MAE loss [%]	Exec Time [ms]
Model 1					
`Float32`	8 607 185	98.47	525.32	+13.18	625.13
`Int8`	8 533 681	30.67	134.85	+18.99	411.16
`Int8` + Float fallback	8 543 683	45.32	134.85	+18.35	411.89
Q-aware `Int8`	8 533 681	30.67	134.85	+1.27	411.46
Q-aware + Float I/O	8 543 683	45.32	134.85	+0.62	412.50
Model 2					
Q-aware `Int8`	4 031 881	30.35	77.44	+2.20	240.30
Q-aware + Float I/O	4 041 883	45.01	77.44	+1.88	240.77

4.2 Real–Field Validation for AE Localization

In this Section, the results in terms of localization accuracy for varying SNR are presented and summarized in Table 2. Each entry expresses the Euclidean distance between the target and the predicted AE source position, by averaging among all the test signals and AE actuation points. As a primary observation, it is worth noticing that the performance of the Q–aware models are absolutely

comparable with the ones obtainable when running the same algorithm on the Keras environment rather than on the edge MCU. This is proven by a maximum deterioration of less than 25 mm in the localization accuracy. The performance is less precise for the model with Float I/O, since it is affected by a more consistent error which increases up to 16 cm in the very unfavourable case of SNR = 4 dB.

More importantly, it is worth stressing that these performances are comparable with the ones reported in [11], where results are presented for Model 1 without quantization running in Keras on a remote PC. Compared to the full–bit version running offline, the edge–enabled variants score a degradation of less than 2 cm, but with much a shorter execution time and model complexity. Such outcome proves the feasibility of the explored TinyML approach for AE signal characterization and paves the way for the development of a novel generation of smart sensors and intelligent systems capable of performing sensor–near data processing and structural inference.

Table 2. Localization accuracy (in cm) of the quantized CNN models while running on the edge processor as a function of decreasing SNR.

SNR [dB]	Model/Quantization Type			
	Model 1 Keras	Model 2 Keras	Model 2 Q-aware `Int8`	Model 2 Q-aware + Float I/O
∞	8.09	8.57	8.67	10.21
20	8.34	8.97	8.84	16.08
16	9.07	8.50	8.75	11.15
12	9.31	9.09	9.32	11.94
8	10.86	10.43	10.66	14.03
4	14.92	16.93	17.01	16.08

5 Conclusions

In this work, the edge implementation of CNN models for the estimation of the ToA in acoustic signals has been attained by pursuing different quantization techniques aimed at shrinking the memory and computational complexity of the network. A resource–constrained device equipped with an STM32F4 MCU has been exploited for prototyping purposes and validated on both synthetic signals and real–field data under different noisy conditions. The obtained outcome shows promising accuracy for the localization of acoustic sources in structural inspections. Therefore, the achieved results discloses new potential for the edge or extreme–edge inference of the structural integrity conditions, thus alleviating the burdensome requirement of transmitting long time series to central processing units while preserving the accuracy of the integrity evaluation process.

References

1. Barat, V., Borodin, Y., Kuzmin, A.: Intelligent ae signal filtering methods. J. Acoust. Emission **28**01, 109–119 (2010)
2. Barat, V., Borodin, Y., Kuzmin, A: Intelligent AE signal filtering methods. J. Acoust. Emission **28** (2010)
3. Bogomolov, D., Testoni, N., Zonzini, F., Malatesta, M., de Marchi, L., Marzani, A.: Acoustic emission structural monitoring through low-cost sensor nodes. In: 10th International Conference on Structural Health Monitoring of Intelligent Infrastructure (2021)
4. Gopinath, S., Ghanathe, N., Seshadri, V., Sharma, R.: Compiling kb-sized machine learning models to tiny Iot devices. In: Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, pp. 79–95 (2019)
5. Jiang, Y., Xu, F.: Research on source location from acoustic emission tomography. In: 30th European Conference on Acoustic Emission Testing & 7th International Conference on Acoustic Emission, Granada, Spain (2012)
6. Pereira, G.: In: Schweiger, G. (ed.) Poverty, Inequality and the Critical Theory of Recognition. vol. 3, pp. 83–106. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45795-2_4
7. Nair, A., Cai, C.S.: Acoustic emission monitoring of bridges: review and case studies. Eng. Struct. **32**(6), 1704–1714 (2010)
8. Pedersen, J.P., Schlanbusch, R., JJ Meyer, T., Caspers, L.W., Shanbhag, V.V.: Acoustic emission-based condition monitoring and remaining useful life prediction of hydraulic cylinder rod seals. Sensors **21**(18) 6012 (2021)
9. ST Microelectronics. UM1472 Discovery kit with STM32F407VG MCU 10 (2020)
10. Warden, P., Situnayake, D.: Tinyml: Machine learning with tensorflow lite on arduino and ultra-low-power microcontrollers. O'Reilly Media (2019)
11. Zonzini, F., Bogomolov, D., Dhamija, T., Testoni, N., De Marchi, L., Marzani, A.: Deep learning approaches for robust time of arrival estimation in acoustic emission monitoring. Sensors **22**(3), 1091 (2022)
12. Zonzini, F., Malatesta, M.M., Bogomolov, D., Testoni, N., Marzani, A., De Marchi, L.: Vibration-based SHM with upscalable and low-cost sensor networks. IEEE Trans. Instrum. Meas. **69**(10), 7990–7998 (2020)

Integrating Electronic Components, Sensors and Actuators in Cast Metal Components: An Overview of the State of the Art

Dirk Lehmhus(✉), Thomas Rahn, Christoph Pille, and Matthias Busse

Fraunhofer Institute for Manufacturing Technology and Advanced Materials (IFAM), Wiener Straße 12, 28359 Bremen, Germany
dirk.lehmhus@ifam.fraunhofer.de

Abstract. Despite the fact that specifically in the automotive industry, several safety-relevant components are produced by casting processes, integrating sensors and electronic components in castings has so far received less attention than the production of smart, self-monitoring components made e.g. from fibre-reinforced plastics (FRP), or by means of additive manufacturing techniques. One reason for this apparent scarcity are the harsh conditions to be endured by integrated systems in processes that rely on processing large amounts of metal in the liquid state. Despite such obstacles, a deeper scrutiny of the topic reveals several studies which solve the underlying problem in quite different ways and in view of different application scenarios such as part identification, load monitoring, damage detection or structural health monitoring. The latter aspect, which has received considerable attention in the aerospace industry in the past, is starting to find interest in the automotive sector, too. A major motivation behind this paradigm shift are autonomous drive and car sharing scenarios, which effectively detach the individual driver from its role in observing the state of the vehicle. The present article summarizes the state of the art in the field, discussing potential applications of smart castings, looking at the various casting processes from lost foam, sand and gravity die casting to low and high pressure die casting, as well as looking at the different sensor and electronic systems being studied in this context.

Keywords: Metal casting · Sensor integration · Smart parts · Material-integrated intelligent systems · Structural health monitoring

1 Introduction

1.1 Application Scenarios

Material-integrated intelligent systems are a growing field of research with several different use cases [1–5]: Part identification e.g. via RFID systems to support production processes or protect against plagiarism as well as the supervision of the actual state of a component during its service life, the latter ranging from mere damage detection to load and structural health monitoring employed to facilitate predictive maintenance, or

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 350–361, 2023.
https://doi.org/10.1007/978-3-031-16281-7_34

support system safety [6], represent the most common of these. Besides, active systems are being considered e.g. for vibration attenuation. These are potentially even more complex, as influencing the behavior of a system must include capturing its state, which makes the monitoring (sensing) task an additional necessity. All of the above applications find their reflection in the foundry industry, too. Apart from mere identification, an important driver for sensory capabilities is the automotive industry as main customer of the light alloy casting sector: The current trends towards autonomous drive as well as car sharing rather than ownership will naturally shift the task of supervising the state of the car and its momentary level of safety from the driver to the car itself. To some degree, this task may be accomplished by the sensor suite anyway available in the car to observe acceleration etc. In combination with a suitable model of the car's structural behavior, information on potential overload can in principle be derived. However, enabling the safety-relevant components to feel their state directly will provide significantly more detail and accuracy, and avoid the problem that once damage has occurred, the evaluation model may lose predictive power due to the change in system state. At the same time, integrated sensors can become a building block for smart manufacturing by providing deeper insight into production processes on a part-by-part basis, or by feeding usage data back into evolutionary development processes [7]. Likely for the first time suggested by Altimus et al. in 1998 as well as Busse et al. in 2006 [8–10], smart castings have since seen a development that justifies the overview attempted with the present text: By now, almost all casting processes as well as a variety of materials and types of sensors have been studied in this context.

1.2 Metal Casting Processes

By definition, casting is a primary shaping process based on metals being processed in the liquid, i. e. molten, state. Some variants exist, though, which process the metal at slightly lower temperatures in the semi-solid state. This implies a double difficulty when it comes to integrating functional systems: for one thing, the temperatures that need to be endured are exceedingly high – Table 1 provides an overview based on [11]. Besides this, contact with the liquid metal implies a high heat transfer coefficient, which means that the thermal energy from the melt is transferred with great efficiency into any kind of embedded structure.

However, at closer scrutiny, boundary conditions do depend significantly on the casting process, too. Figure 1 further below provides an overview of the most common casting processes – the typical distinction used here, too, is based on the nature of the mold and the model or pattern, the latter being used in many processes to create the former: Both can either be permanent or lost, i.e. destroyed in the course of part production. Best known lost pattern process is investment or precision casting, which uses a positive wax model of the final part. This model is enclosed by a molding material and removed by melting, freeing the cavity for the actual casting process – hence the process is also known as lost wax casting. On removing the part from this mold, the latter is destroyed just like the pattern. Sand casting in contrast uses a lost mold made of organically or inorganically bonded grains of sand. Production of this mold often relies on permanent models, but nowadays, direct printing using a digital model is common, too. In both processes, the mold is typically heated up before casting, and heat extraction is slow

also due to limited thermal conductivity of molding materials. In contrast, permanent metallic molds as used in gravity die casting (GDC), low pressure die casting (LPDC) or high pressure die casting (HPDC) allow faster extraction of heat, with typical cooling rates increasing in the order given above if similar part volumes and surface-to-volume ratios are compared. In HPDC, cooling rates of 100 K/s may thus be achieved. Beyond type of mold and model, pressure assistance to support feeding is a distinguishing feature of processes. Here, LPDC and HPDC are the most important process variants: While in the former, pressure levels are typically below one bar, in the latter case, during the intensification phase levels beyond 1000 bar (100 MPa) can be reached.

Table 1. Typical casting temperatures for different alloy systems according to [11].

Alloy system	Casting temperature lower limit [°C]	Upper limit [°C]
Zn and Zn alloys	420	580
Al and Al alloys	620	730
Mg and Mg alloys	620	730
Cu and Cu alloys	960 (Cu: 1150)	1450 (Cu: 1290)
Cast iron (low-alloy)	1340	1480
Steel (no to low-alloy)	1560	1700
Steel (high-alloy)	1500	1640
Ti and Ti alloys	1700	1800

The primary factors that control the suitability of a specific casting process for smart system integration thus are the following:

- temperature level and heat capacity (primarily alloy-dependent)
- temperature exposure time or heat input (process-, but also part-dependent)
- pressure level (process-dependent)
- melt flow velocity (process-dependent)

While this holds true on a general level, the part-dependent situation can influence the process outcome, too. Superimposed aspects include part geometry as well as positioning of the system to be integrated (e.g. relative to the melt flow direction, near part surface or completely embedded, etc.).

It is important to understand that adversary effects of these boundary conditions can differ significantly. While destruction of the system is one possible issue, loss or modification of functional properties is another: For example, piezoelectric as well as other ferroic materials will lose polarization and thus functionality when transgressing their Curie temperature. The most common piezoelectric material, PZT, exhibits Curie temperatures between 160 and 365 °C, roughly [12]. Limitations may also apply for solders, for passives like resistors or capacitors, for energy storage devices and of course for the electronics needed for data evaluation, too. Such high temperature electronics

(HTE) research has recently been summarized by Watson and Castro [13] as well as Amalu et al. [14] for electronics, Dziedzic et al. for passives [15], Bove et al. and Kazys for piezoelectric devices [12, 16] and Zhang as well as Zhou et al. for packaging solutions [17–19]. According to Watson and Castro, the critical realm in most cases, and thus the limitation of HTE as a research and application area, starts from 150 °C upwards [13].

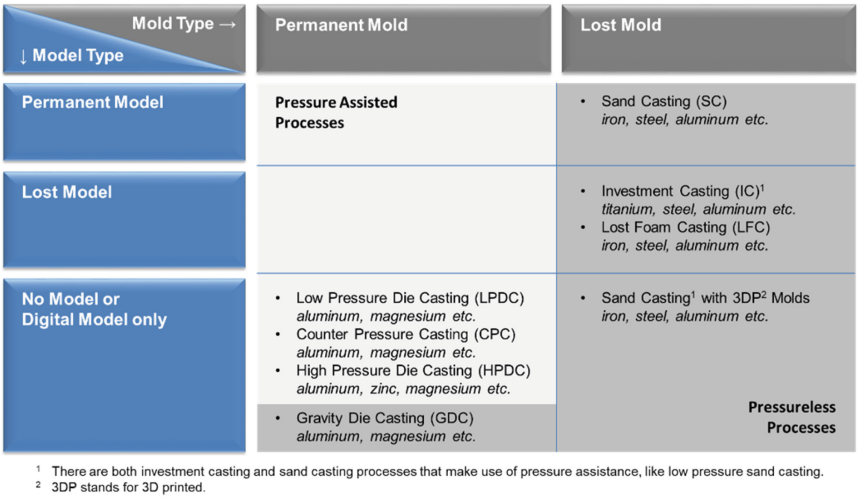


Fig. 1. Distinction of casting processes based on the nature of mold and model or pattern. For each process, common alloy systems are named.

Having said this, in principle, there are four fundamental ways of helping functional systems through the hardships of the casting process: **harden** the system itself, **protect** it by mean of an encapsulation, **simplify** it, or **distribute** it in such a way that most vulnerable elements can be placed outside the casting and thus away from any contact with the liquid metal. Needless to say, in many cases combinations of the respective techniques are employed. In the following section, several sensor integration case studies are presented and discussed in terms of the approach chosen. A distinction is made between pressureless (SC, LFC, GDC) and pressure-assisted processes (LPDC, HPDC). In the latter case, despite considerable interest specifically in structural monitoring of high quality LPDC components, no published example could be identified, meaning that the respective chapter will focus on the HPDC process.

2 Case Studies

2.1 Pressureless Casting Processes: Sand Casting and Gravity Die Casting

Sensor Integration Use Cases. As stated above, sand and gravity die casting processes stand for highest thermal loads on any insert due to the lower cooling rates, which are based on comparatively limited heat extraction through the mold (obvious specifically for

sand casting) in conjunction with higher wall thicknesses and thus larger melt volume and lower surface-to-volume ratio than e.g. in HPDC. This is why tuning sensor concepts to survive requires specific effort in this case. As a consequence, distribution of components and simplification of sensors gain importance.

Probably the easiest way of detecting damage is a binary one – a rip wire approach: the idea is to integrate a conductive path in the respective component and transmit some signal – usually an electrical one – through it. In case of failure, signal transmission is interrupted. This method has been realised in GDC components. In this case, the rip wire was led through ceramic tubes meant to provide electrical insulation against the casting. The principle is depicted in Fig. 2 (a). A major challenge associated with this approach is to make sure that (a) the overload signal is obtained prior to complete, catastrophic failure of the part and (b) fracture does not bypass the wire. The latter could happen if the wire's ductility exceeds that of both casting and insulating material.

A piezoresistive sensor system for casting integration has been suggested by Tiedemann et al. based on the Heraeus Celcion® thick film materials for screen printing, which requires sintering at temperatures above 600 °C [20]. In their study, Tiedemann et al. applied sensors and conductive paths on an AlMg3 substrate and cast an AlSi7Mg0.3 aluminum alloy around it. The temperature response of the sensors was tested before and after casting as a proof of functionality. A linear relationship between temperature and relative change of resistance was observed in the range from RT to 80 °C, though sensitivity of both sensor designs decreased after casting [21].

Most functional systems considered for smart products rely on electrical signals. This adds the difficulty of ensuring electrical insulation for system and any substrate on which it is built. This difficulty can be avoided if light is used as signal carrying medium. Typical example of such an approach are fiber-optic sensors e. g. based on Fibre-Bragg-Gratings (FBG). The measurement principle is wavelength-dependent reflection of light by a grating integrated in the optical fiber via a local variation of the refractive index with defined periodicity, which determines the wavelength reflected in the unloaded reference state. Straining of the fiber affords a shift in the reflected wavelength. Selectivity of each grating allows multiple sensor spots to be inscribed over the length of a single optical fiber, provided they are tuned to a different wavelengths. In a casting context, the technique has extensively been studied at the Technical University of Munich (TUM), originally by Weraneck et al. with the aim of monitoring solidification processes in gravity die casting [22, 23]. The problem of stability of inscribed gratings is circumvented by the use of so-called regenerated FBGs or RFBGs – in these, a seed grating is initially lost, then reformed at elevated temperatures to afterwards sustain temperatures of up to and above 1000 °C [24, 25]. In a logical extension of the original work by Weraneck et al., Lindner et al. concentrated on the response of the integrated sensor to thermal and mechanical loads, thus moving towards condition monitoring applications [26, 27]. The same approach has since been chosen by Bian et al. for gravity die casting of a CuSn2 alloy at >1100 °C [25].

An alternative concept based on the "simplify" approach has been evaluated by Carlsson et al.: here, simple, wire-type sensors are combined with uncommon sensing principles which relinquish the need for electrical insulation - instead of electrical or optical signals, mechanical oscillations are introduced at one end of the embedded wire.

Amplitude and frequency shift of the transmitted signal captured at the wire's opposite end correspond to loads acting on the host component. Here, complexity is shifted to signal conditioning and analysis, facilitating an extremely robust sensor design. The principle and its capabilities with respect to detecting mechanical has since been successfully demonstrated in sand casting both of aluminum and iron components and thus represents one of the few concepts validated for the latter, specifically demanding application [28].

2.2 Pressure-Assisted Casting Processes: High Pressure Die Casting

Part Identification Scenarios. Part identification is conventionally being done e.g. by labelling, laser marking, needle printing or embossing. All these techniques modify the part surface and are thus prone to corruption or manipulation. For this reason, there is an interest in alternatives. RFIDs as standalone systems requiring neither interconnects nor internal power supply for readout are comparatively easy to integrate: in metal components, electromagnetic shielding has to be overcome, and they don't lend themselves to distribution approaches, but decoupling from the host component via thermal protection measures is less of a problem than in the case of sensors.

Pille et al. have reported about RFID systems integrated in HPDC components as early as 2010: within the framework of the Collaborative Research Centre (CRC) SFB 637 Dynamics in Logistic Systems funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), RFID-based part identification allowed the components passing through a model production line to actively control their path and thus their individual processing and assembly [29, 30]. A further development of this basic concept towards higher technology readiness levels (TRL) was realized in the

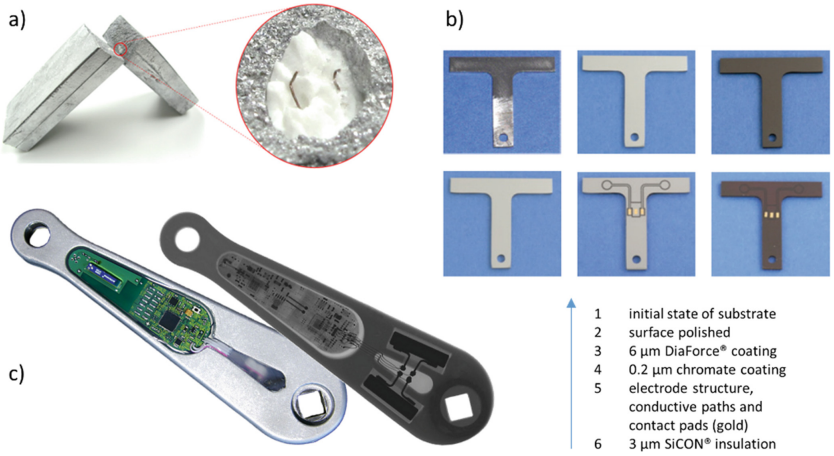


Fig. 2. Examples of casting-integrated sensors: (a), rip wire approach realized in GDC of aluminum alloy, (b) processing steps of a piezoresistive sensor based on the DiaForce® technology developed by Fraunhofer IST [32–34] for HPDC of aluminum alloy, (c) piezoelectric transducer in a pedal crank arm with associated electronics added after HPDC casting of aluminum alloy (all images by Fraunhofer IFAM) [35].

course of the European project MUSIC (MUlti-layer control & cognitive System to drive metal and plastic production line for Injected Components, Collaborative IP, call FP7-FoF-ICT-2011.7.1, proposal number 314145), which included incorporating the pick-and-place operation for positioning the sensor within the mold prior to casting in a HPDC process chain by making use of an available robotic spraying system for manipulation [31]. Generally, a protect scheme was applied with the RFID system being encapsulated e.g. by a polymer which would at the same time provide a dielectric slot for read-out.

Sensor Integration Use Cases. As has already been shown in the preceding section, thanks to fast cooling rates, HPDC processes allow consideration of more complex solutions for integration. An example in this respect are piezoresistive sensors, best known in the form of strain gauges, which make use of changes in electrical resistance induced by mechanical strain. Among these are thin film sensors as produced from diamond like carbon materials (DLC, DiaForce®, [32, 33]). Both the DLC sensor, the conductive paths with contact pads and the insulation layers (SiCON®, [36]) exhibit high temperature stability up to and beyond the casting temperature. The system thus constitutes an example of the "harden" approach. The sensors and contacts were built up on a polished steel substrate by means of PAVD processes and integrated in a HPDC component. In subsequent cyclic bending tests, the sensor function could successfully be demonstrated [37]. In a separate study, the same screen printing materials (Heraeus Celcion®) already evaluated in GDC were also tested in HPDC [38].

Also unlike RFID systems, sensors typically require peripherals like energy supply, communications, interconnects and potentially some computational power. As a consequence, "distribute" solutions gain importance, limiting the number of functionalities actually integrated to sensors and actuators, while data evaluation, power supply etc. are provided externally. An example of such a system configuration has already been presented in Fig. 2 (c). In this case, two piezoelectric transducers were integrated in a pedal crank arm for force measurement. These active components were shielded against the thermal impact of the casting process, as conventional PZT instead of high temperature piezoelectrics was used and the metallic encasement would have prevented or at least impeded (re-) polarization of the material [35]. A general issue regarding most PZT materials is their brittle nature, which may also warrant protection, e.g. in view of thermally induced stresses. Beyond force measurement, the piezoelectric elements were also evaluated successfully as energy harvesting devices. In a similar approach, though using temperature differences rather than cyclic mechanical strain, thermoelectric energy harvestors or thermogenerators were integrated in HPDC components by Ibragimov et al., who explicitly built up their devices from high temperature materials (borosilicate glass substrate, thermocouples based on WTi + Pt and Si single crystal, SiO_2 and Al_2O_3 insulation layers), favouring the "harden" over the "protect" strategy. The choice of the material combination for the thermocouples meant that efficiency was sacrificed in exchange for high-temperature stability not offered by conventional thermoelectrics like bismuth tellurides [38, 39].

Actuation and Vibration Attenuation Use Cases. Used for sensing of dynamic strains, piezoelectrics can also be employed as actuators to suppress such strains.

Friedrich-Alexander University of Erlangen has a long development history in incorporating piezoelectric transducers in HPDC parts for vibration control. Initial work by Rübner et al. used commercial d_{31} patch transducer made from monolithic PZT plates contacted by a flexible copper mesh on either side. Insulation of transducer, mesh and contact wires relied on a polyimide casing, a high temperature thermoplast. A so-called expanded metal aluminum wire mesh-like structure enveloped the active element and provided structural support against mechanical forces without impeding melt infiltration, avoiding relocation by the incoming melt (which reached up to 100 m/s in these experiments). Initial studies focused on the sensory function, which was successfully demonstrated by recording the voltage signal from the PZT transducer during free vibration of the casting [40, 41]. Parallel work by Flössel et al. used low temperature co-fired ceramics (LTCC) as insulating layer [42]. Klassen et al. continued work on polyimide-encased systems, investigating the influence of HPDC process variation on PZT stability. Asymmetric, surface-near integration of devices was observed to induce formation of cracks as well as perforations within the active element. Via casting simulation, these were linked to filling patterns and the resulting inhomogeneous temperature distribution, as well as mechanical forces exerted by the melt [43]. Active performance of the piezoelectric modules was scrutinized by Schwankl et al., showing almost no loss in thin-, but some degradation in thick-walled regions [44]. Frequency-sweep studies by Schwankl et al. confirmed mode shape and resonance frequency simulations, showing that the respective modes could indeed be excited, while providing the excitation signal to the PZT transducer with an adapted phase shift allowed for vibration attenuation [45]. In a concluding work, Stein et al. described a full process chain for industrial-scale equipment of HPDC parts with LTCC-encapsulated PZT transducers [46].

3 Summary and Conclusion

There are good reasons for integrating electronic systems into castings, and several viable approaches exist, as the present overview has shown. Nowadays, these cover most casting processes, ranging from sand to high pressure die casting, and several materials, including cast iron at the high temperature end of the scale. Aluminum, however, clearly remains the dominant material when it comes to the number of studies published. While initial developments naturally focused on tasks easier to realize, we are now seeing trends towards the harder tasks which are partly dictated by application needs: a prominent example is the implementation of structural health monitoring in aluminum components produced by means of LPDC, as this is the high quality process typically employed for safety-relevant components in the automotive area, which warrant constant supervision of state most. Realization is further impeded in this case by the fact that such castings are often subjected to T6 heat treatments which typically involve holding the part above 500 °C for several hours during the solution heat treatment step.

The paths towards realizing these applications can be manifold and follow either of the strategies described above via the catchphrases protect, harden, simplify and distribute, or combinations of these. Additive Manufacturing (AM), which is sometimes seen as a potential competitor to casting in the mid to long term perspective, may prove

to be an ally here: smart system integration is an area where strong synergies between both manufacturing processes exist, as AM can be employed to encapsulate a system, too [47–50]. Furthermore, gaining in importance is the re-evaluation of sensing principles: while many authors focus on well-known piezoelectricity or –resistivity, alternatives are starting to emerge, ranging from fiber-optic sensors to solutions based on mechanical oscillations – and there may be further options out there. Such fundamentally different approaches may provide alternatives for realizing systems employing readily available materials with high thermal stability instead of having to rely on advanced and cost-intensive HTE materials and processes.

In terms of application scenarios, it will be interesting to observe which new ideas may surface with respect to the use of sensor data already during manufacturing, e. g. as a means to gather data for feeding in to digital twins of the respective processes.

Despite or rather because of such promising perspectives, the research area as such remains a treasure trove of opportunity, as many of the studies published so far concentrate on ensuring survivability of systems, while e.g. performance of the resulting components has not been scrutinized accordingly yet. Having said this, the most important step for the technologies presented here has not been taken yet: Commercialization, i. e. the introduction of sensors in series production parts made through any of the above casting processes is, to the authors' best knowledge, still pending.

References

1. Lang, W., et al.: From embedded sensors to sensorial materials – the road to function scale integration. Sens. Actuators, A **171**(1), 3–11 (2011)
2. McEvoy, M.A., Correll, N.: Materials that combine sensing, actuation, computation and communication. Science **347**, 1261689 (2015)
3. Mekid, S., Saheb, N., Khan, S.M.A., Qureshi, K.K.: Towards sensor array materials: can failure be delayed? Sci. Technol. Adv. Mater. **16**, 034607 (2015)
4. Bosse, S., Lehmhus, D., Lang, W., Busse, M. (eds.): Material-Integrated Intelligent Systems – Technology and Applications. Wiley, Hoboken (2018)
5. Bosse, S., Lehmhus, D.: On concepts and challenges of realizing material-integrated intelligent systems. In: Bosse, S., Lehmhus, D., Lang, W., Busse, M. (eds.) Material-Integrated Intelligent Systems – Technology and Applications, pp. 1–28. Wiley, Hoboken (2018)
6. Lehmhus, D., Busse, M.: Structural health monitoring. In: Bosse, S., Lehmhus, D., Lang, W., Busse, M. (eds.) Material-Integrated Intelligent Systems – Technology and Applications, pp. 529–570. Wiley, Hoboken (2018)
7. Lehmhus, D., et al.: Cloud-based automated design and additive manufacturing: a usage data-enabled paradigm shift. Sensors **15**, 32079 (2015)
8. Altimus, J.C., Johnson, V.D.: Remote identification of metal castings. Trans. Am. Foundrymen's Soc. **106**, 605–608 (1998)
9. Busse, M., Wöstmann, F.-J., Müller, T., Melz, T., Spies, P.: Intelligente Gussteile - Einsatz adaptronischer Komponenten in Kombination mit Gussteilen. Giesserei **93**, 48–53 (2006)
10. Busse, M., Wöstmann, F.-J., Müller, T., Melz, T., Spies, P.: Intelligente Druckgussteile. Metall **60**, 738–741 (2006)
11. https://www.giesserei-praxis.de/giesserei-lexikon/glossar/giesstemperatur. Accessed 4 May 2022
12. Kazys, R., Vaskeliene, V.: High temperature ultrasonic transducers: a review. Sensors **21**(9), 3200 (2021)

13. Watson, J., Castro, G.: A review of high-temperature electronics technology and applications. J. Mater. Sci.: Mater. Electron. **26**(12), 9226–9235 (2015). https://doi.org/10.1007/s10854-015-3459-4
14. Amalu, E.H., Ekere, N.N., Bhatti, R.: High temperature electronics: R&D challenges and trends in materials, packaging and interconnection technology. In: Proceedings of the 2nd International Conference on Adaptive Science & Technology (ICAST 2009), Accra, Ghana, 14–16 January 2009. https://doi.org/10.1109/ICASTECH.2009.5409731
15. Dziedzic, A., Nowak, D.: Thick-film and LTCC passive components for high-temperature electronics. Radioengineering **22**, 218–226 (2013)
16. Bove, T., Damjanovic, D., Liang, K., Wolny, W.: Piezoceramic materials for high-temperature & high-pressure applications in oilfield exploration & production. In: Proceedings of the 2013 Joint IEEE International Symposium on Applications of Ferroelectric and Workshop on Piezoresponse Force Microscopy (ISAF/PFM), Prague, Czech Republic, 21–25 July 2013, pp. 62–65. https://doi.org/10.1109/ISAF.2013.6748669
17. Zhang, H., Minter, J., Lee, N.-C.: A brief review on high-temperature, Pb-free die-attach materials. J. Electron. Mater. **48**, 201–210 (2019)
18. Zhou, Z., Cui, J., Yu, F., Johnson, R.W., Hamilton, M.C.: Evaluation of thick-film materials for high-temperature packaging. IEEE Trans. Compon. Packag. Manuf. Technol. **8**, 773–783 (2018)
19. Zhou, Z., Johnson, R.W., Hamilton, M.C.: Mechanical reliability of thick films for high-temperature packaging. IEEE Trans. Compon. Packag. Manuf. Technol. **8**, 1003–1013 (2018)
20. Heraeus Product Brochure. Heraeus Celcion—Materials System for LED Circuits. https://www.heraeus.com/media/media/het/doc_het/products_and_solutions_het_documents/thick_film/Brochure_Heraeus_Celcion_-_Materials_System_for_LED_Circuits.pdf. Accessed 4 May 2022
21. Tiedemann, R., Fischer, M., Busse, M., Lang, W.: Integrating sensors into castings made of aluminum - new approaches for direct sensor integration in gravity die casting. Procedia Manuf. **24**, 179–184 (2018). https://doi.org/10.1016/j.promfg.2018.06.042
22. Weraneck, K., et al.: Strain measurement in aluminium alloy during the solidification process using embedded fibre Bragg gratings. Sensors **16**, 1853 (2016)
23. Lindner, M., et al.: Regenerated Bragg grating sensor array for temperature measurements during an aluminum casting process. IEEE Sens. J. **18**, 5352–5360 (2018)
24. Polz, L., Dutz, F.J., Maier, R.R.J., Bartelt, H., Roths, J.: Regenerated fibre Bragg gratings: a critical assessment of more than 20 years of investigations. Opt. Laser Technol. **134**, 106650 (2021)
25. Bian, Q., Bauer, C., Stadler, A., Jakobi, A., Koch, A.W., Roths, J.: Multipoint temperature monitoring based on a regenerated fiber bragg grating temperature sensor array in copper casting. In: Proceedings of SPIE 11591, Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2021, 115910U, 22 March 2021. https://doi.org/10.1117/12.2588600
26. Heilmeier, F., et al.: In-situ strain measurements in the plastic deformation regime inside casted parts using fibre-optical strain sensors. Prod. Eng. Res. Devel. **13**(3–4), 351–360 (2019). https://doi.org/10.1007/s11740-019-00874-7
27. Lindner, M., et al.: Fiber Bragg sensors embedded in cast aluminum parts: axial strain and temperature response. Sensors **21**, 1680 (2021)
28. Carlsson, R., et al.: Sensors integrated inside metal castings verified to respond to force. In: Proceedings of the 9th ECCOMAS Thematic Conference on Smart Structures and Materials (SMART 2019), Paris, France, 8–11 July 2019
29. Pille, C.: In-process embedding of piezo sensors and RFID transponders into cast parts for autonomous manufacturing logistics. In: Proceedings of Smart Systems Integration 2010, Como, Italy, 23–24 March 2010

30. Hribernik, K.A., Pille, C., Jeken, O., Thoben, K.-D., Windt, K., Busse, M.: Autonomous control of intelligent products in beginning of life processes. In: Proceedings of the International Conference on Product Life Cycle Management, Bremen, Germany, 12–14 July 2010
31. Bonollo, F., Gramegna, N.: The MUSIC guide to the key-parameters in High Pressure Die Casting. Assomet servizi srl, Enginsoft SpA (2014). ISBN 978-8887786-10-1
32. Peiner, E., Tibrewala, A., Lüthje, H., Bandorf, R., Biehl, S., Doering, L.: Piezoresistive diamond-like carbon micro strain gauges. In: Proceedings of the XVIII IMEKO World Congress – Metrology for a Sustainable Development; Rio de Janeiro, Brazil, 17–22 September 2006
33. Bräuer, G., Bandorf, R., Biehl, S., Dietz, A., Lüthje, H., Vergöhl, M.: Smart coatings for intelligent surfaces. VIP Vacuum's Best **20**(S1), 34–37 (2008)
34. Pille, C., Biehl, S., Busse, M.: Encapsulating piezoresistive thin film sensors based on amorphous diamond-like carbon in aluminium castings. In: Proceedings of the 1st Joint International Symposium on System Integrated Intelligence, Hanover, Germany, 27–29 June 2012, pp. 192–194 (2012)
35. Tiedemann, R., Pille, C., Dumstorff, G., Lang, W.: Sensor integration in castings made of aluminum - new approaches for direct sensor integration in aluminum high pressure die casting. Key Eng. Mater. **742**, 786–792 (2017)
36. Mayer, D., Melz, T., Pille, C., Woestmann, F.-J.: CASTRONICS - direct integration of piezo ceramic materials in high pressure die casting parts for vibration control. In Proceedings of Actuator 2008, 11th International Conference on New Actuators & 5th International Exhibition on Smart Actuators and Drive Systems, Bremen, Germany, 9–11 June 2008
37. Bewilogua, K., Bialuch, I., Ruske, H., Weigel, K.: Preparation of a-C:H/a-C:H:Si: O and a-C:H/a-C:H: Si multilayer coatings by PACVD. Surf. Coat. Technol. **206**, 623–629 (2011)
38. Ibragimov, A., Pleteit, H., Pille, C., Lang, W.: Micromachined thermogenerator directly integrated into metal parts: technological aspects of the embedding process. In: Proceedings of the 1st Joint International Symposium on System-Integrated Intelligence, Hanover, Germany, 27–29 June 2012, pp. 204–207 (2012)
39. Ibragimov, A., et al.: A thermoelectric energy harvester directly embedded into casted aluminum. IEEE Electron Device Lett. **33**(2), 233–235 (2012)
40. Bräutigam, V., Schellack, T., Körner, C., Singer, R.F.: Optimized fixing technique for the integration of piezoceramic modules in die castings. Giessereiforschung **58**, 11–15 (2006)
41. Rübner, M., Körner, C., Singer, R.F.: Integration of piezoceramic modules into die castings – procedure and functionalities. In: Advances in Science and Technology, vol. 56, pp. 170–175 (2008)
42. Flössel, M., Scheithauer, U., Gebhardt, S., Schönecker, A., Michaelis, A.: Robust LTCC/PZT sensor-actuator module for aluminium die casting. In: Proceedings of the 2009 European Microelectronics and Packaging Conference, Rimini, Italy, 15–18 June 2009. https://ieeexplore.ieee.org/abstract/document/5272855
43. Klassen, A., et al.: Influence of the fabrication process on the functionality of piezoceramic patch transducers embedded in aluminum die castings. Smart Mater. Struct. **21**, 115014 (2012)
44. Schwankl, M., et al.: Active functionality of piezoceramic modules integrated in aluminum high pressure die castings. Sens. Actuators, A **207**, 84–90 (2014)
45. Schwankl, M., Kimme, S., Pohle, C., Drossel, W.-G., Körner, C.: Active vibration damping in structural aluminum die castings via piezoelectricity – technology and characterization. Adv. Eng. Mater. **17**, 969–975 (2015)
46. Stein, S., et al.: A process chain for integrating piezoelectric transducers into aluminum die castings to generate smart lightweight structures. Results Phys. **7**, 2534–2539 (2017)
47. Isaza-Paz, J., Wilbig, J., Aumund-Kopp, C., Petzoldt, F.: RFID transponder integration in metal surgical instruments produced by additive manufacturing. Powder Metall. **57**, 365–372 (2014)

48. Lehmhus, D., et al.: Customized smartness: a survey on links between additive manufacturing and sensor integration. Procedia Technol. **26**, 284–301 (2016)
49. Hehr, A., et al.: Integrating fiber optic strain sensors into metal using ultrasonic additive manufacturing. JOM **70**(3), 315–320 (2017). https://doi.org/10.1007/s11837-017-2709-8
50. Monaghan, T., Capel, A.J., Christie, S.D., Harris, R.A., Friel, R.J.: Solid-state additive manufacturing for metallized optical fiber integration. Compos. Part A **76**, 181–193 (2015)

Fast Temperature-Compensated Method for Damage Detection and Structural Health Monitoring with Guided Ultrasonic Waves and Embedded Systems

Christoph Polle[1(✉)], Stefan Bosse[2,3], Michael Koerdt[1], Björn Maack[1], and Axel S. Herrmann[1]

[1] Faserinstitut Bremen e.V (FIBRE), Am Biologischen Garten 2, 28359 Bremen, Germany
polle@faserinstitut.de

[2] Department Mathematics and Computer Science, University of Bremen, 28359 Bremen, Germany

[3] Institute for Digitization, 28359 Bremen, Germany
http://www.faserinstitut.de/

Abstract. In this paper, the temperature scaling approach (TS) for temperature compensation of guided ultrasonic waves (GUW) is presented. In many previous works, it could be observed that the amplitudes of the GUW change with temperature. In the TS approach, it is assumed that for each transducer configuration several functions $M(f,T)$ exist that describe the change of the maximum amplitude depending on the frequency f and the temperature T. It is also assumed that several functions $F(f,T)$ exist that describe the changes of the full width at half maximum (FWHM) from the amplitude containing the maximum value of the measurement. Furthermore, it is assumed that the functions $M(f,T)$ and $F(f,T)$ also depend on the medium in which the GUW propagate, so that a change of the medium results in a change of the functions. This means that these functions can be used for damage detection. GUW data from the http://openguidedwaves.de/ database were analyzed with the TS approach. Data from a carbon fiber reinforced polymer plate with a temperature range from 20 °C to 60 °C were analyzed for different actuator-sensor paths and compared with position-dependent artificial defects. To determine the functions $M(f,T)$ and $F(f,T)$, polynomials were fitted to the measured values of the maxima and full FWHM. The TS analysis showed a detect ability of the artificial defects over the entire temperature range.

Keywords: SHM · Temperature compensating · Signal processing · Data analysis · Embedded systems

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 362–378, 2023.
https://doi.org/10.1007/978-3-031-16281-7_35

1 Introduction

With its potential to optimize or eliminate maintenance cycles and the emerging opportunity to use lightweight materials more extensively, structural health monitoring (SHM) has attracted much attention in recent years. Whereas methods with guided ultrasonic waves (GUW) are interesting for many applications due to the possibility of examining a large area for damage with few sensors [1,2]. Especially the integration of a sensor network has the advantage that a large area can be checked quickly by few personnel, as the sensors do not have to be applied manually. In addition, measurement errors are avoided because the sensors remain stationary in an integrated system. However, a major problem in the practical use of SHM systems based on GUW is the temperature effect on the signals [3,4]. There have been many methods developed to address the issue, like optimal baseline selection (OBS) [5], baseline signal stretch (BSS) [6,7] or dynamic time wrapping (DTW) [8] but, these methods are often based on memory intensive baselines and/or computationally intensive adjustments and comparisons [9]. However, if you want to integrate sensor nodes with small electronics such as micro controllers into a material for monitoring, you need compensation methods that are both cheap in computation and memory. An interesting approach to temperature compensation for permanent monitoring of a pipe system that meets these requirements is the location specific temperature compensation (LSTC) by Mariani et al. [10,11]. In LSTC a torsional, T(0,1) GUW mode is sent into the pipe via a transducer ring that is clamped around the pipe. Subsequently, the reflections of the wave from welds and end pieces are detected by the transducer ring. Using the temperature-dependent wave speed, it is now possible to determine points on the time signal which correspond to a specific locations L on the tube. This means that a temperature compensation must be made with respect to the wave speed. Now a function of the amplitude $A(T, L)$ as a function of the temperature can be determined for specific locations L. If damage occurs before the location L, this also has an influence on the amplitude at location L, so that the residual signal between the local amplitude function $A(T, L)$ and the measured amplitude becomes larger, with which the damage is detected. LSTC was also applied on bulk waves [12]. In this paper we present a temperature compensation method for GUW pitch catch measurements. We call our method temperature scaling (TS). TS follows a similar idea as LSTC and is also like LSTC cheap in computing and storage capacity. So that TS can be applied for simple damage detection (is there damage yes/no?) with in micro controllers integrated in sensor nodes. The features extracted from TS have also been successfully used for damage localization [13]. Like the LSTC, the TS uses temperature dependent functions to achieve temperature compensation. As TS is a compensation method for pitch catch measurements, TS does not use the amplitude at a specific location but temperature-dependent features of the entire time signal. Also TS provides under certain conditions the possibility of a temperature compensation without the knowledge of the current temperature.

Beside the analytical derivation of temperature-scaled signal-damage features, the implementation of the algorithms used for damage detection are discussed and evaluated. The computed signal features can be used as input for Machine Learning (ML). Since the target application is a SHM system finally integrated in the material (material-integrated intelligent system [14]), hard resource constraints must be considered. There are efforts to map ML models trained on generic desktop computers on embedded systems (TinyML, a survey can be found in [15]). The Sect. 6 is dedicated to computational and memory requirements of the signal feature computation and damage prediction.

2 Damage Detection with the Temperature Scaling Method

At this point, the temperature scaling (TS) method is presented. The TS approach takes into account that certain features of GUW depend on the temperature of the medium in which the waves propagate like the amplitude of the signal [16]. In order to keep the required computing effort and storage space as low as possible, the TS does not consider the entire amplitude of the signal but only individual features. These are the maximum value of the signal M, the position of M on the time axis $t_{\max}$ and the full width at half maximum F of the peak containing the maximum value. These features are particularly suitable because they are easy to identify in any signal and because they are particularly affected by changes in the GUW propagation medium. This is because M is often composed of different signal parts, such as A and S modes of the dispersive wave or constructive interference with reflected signal parts. So if there is a change in one of the signal parts that contribute to the value of M, the value of M changes and so change the values of $t_{\max}$ and F. In the TS method, it is assumed that the maximum M and F of the envelope of a GUW catch signal, with a certain pitch signal frequency f, do not change randomly with the temperature T. Thus, these changes can be described as functions $M(f,T)$ and $F(f,T)$. Furthermore, it is assumed that these functions are also dependent on the medium in which the GUW propagates. Thus, a change in the medium, e.g. a damage, leads to changes of $M(f,T)$ and $F(f,T)$ [17]. These changes caused by damage can be due to the loss of energy of one or more GUW modes when the actuated wave passes through the damage, due to wave reflections from the damage or due to the fact that the damage site itself starts to oscillate after the actuated wave has passed through and thus emits waves in all directions. The last two effects lead to interference with the actuated waves, which means that a damage that is not directly between the actuator and the sensor can also be detected.

If the functions $M_0(f,T)$ and/or $F_0(f,T)$ for the undamaged case are known for a certain pitch-catch transducer configuration, these functions can be used for damage detection. The lower index 0 at the functions indicates that there

is no damage present. So for a given pitch-catch transducer configuration, with the known functions $M_0(f,T)$ and $F_0(f,T)$, the idea is to compare the measured maximum $M_\mathrm{m}(f_\mathrm{m},T_\mathrm{m})$ of the catch envelope with pitch frequency f_m at temperature T_m with the value of $M_0(f_\mathrm{m},T_\mathrm{m})$. With the following conditions applying:

$$M_\mathrm{m}(f_\mathrm{m},T_\mathrm{m}) = M_0(f_\mathrm{m},T_\mathrm{m}) \pm \Delta M_0 \Rightarrow \text{No damage}, \tag{1}$$

$$M_\mathrm{m}(f_\mathrm{m},T_\mathrm{m}) \gtrless M_0(f_\mathrm{m},T_\mathrm{m}) \pm \Delta M_0 \Rightarrow \text{Damage}. \tag{2}$$

with ΔM_0 being the value by which $M_\mathrm{m}(f_\mathrm{m},T_\mathrm{m})$ is allowed to maximally deviate from $M_0(f_\mathrm{m},T_m)$ in order to pass as damage-free. Similar conditions are defined for the F:

$$F_\mathrm{m}(f_\mathrm{m},T_\mathrm{m}) = F_0(f_\mathrm{m},T_\mathrm{m}) \pm \Delta F_0 \Rightarrow \text{No damage}, \tag{3}$$

$$F_\mathrm{m}(f_\mathrm{m},T_\mathrm{m}) \gtrless F_0(f_\mathrm{m},T_\mathrm{m}) \pm \Delta F_0 \Rightarrow \text{Damage}. \tag{4}$$

3 Test Data

To proof the assumptions made above, data sets [18] of the Open Guided Waves platform were used. All data sets contain GUW pitch-catch measurements recorded in a temperature range from 20 °C to 60 °C with temperature steps of 0.5 °C. At each target temperature, pitch-catch measurements with a narrow band 5-cycle Hann-filtered sine wave pitch signal were performed for core frequencies in a range of 40 kHz to 260 kHz in 20 kHz steps. All measurements were performed on a carbon fiber reinforced plastic (CFRP) plate (500 mm × 500 mm × 2 mm) made from Hexply®M21/34%/UD134/T700/300 carbon pre-impregnated fibers with a stacking of $[45°/0°/-45°/90°/-45°/0°/45°/90°]_S$. For the pitch-catch measurements 12 DuraAct piezoelectric transducers were co-bonded to the plate. To simulate damage [18], four artificial damage were placed at different positions at the surface. The artificial damage consisted of circular aluminum disk with a diameter of 1 cm, which was attached to the CFRP plate with tacky tape. A sketch of the plate setup can be seen in Fig. 1. For a more detailed description of the measurements, please see [16]. In the following, only transducer pairs were considered where the actuator transducer i consists of the set $\{1,2,3,4,5,6\}$ and the sensor transducer j consists of the set $\{7,8,9,10,11,12\}$.

4 Temperature Scaling Data Processing

In the following, we describe how the raw data were processed for the TS method. At first, the analytical signal of the data was computed. For a given time dependent signal $s(t)$ the analytical signal $s_\mathrm{a}(t)$ is given as:

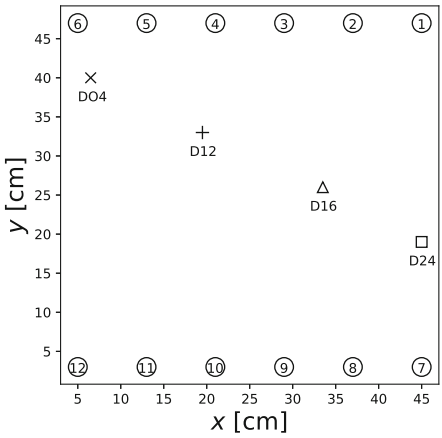


Fig. 1. Sketch of the transducer configuration and artificial damage positions D04 (×), D12 (+), D16 (△) and D24 (□) on the investigated CFRP plate. Circles with numbers represent the transducers. The damage and transducer indexing corresponds to [16] and [18].

$$s_{\mathrm{a}}(t) = s(t) + \mathrm{i}s(t) * \frac{1}{\pi t} \tag{5}$$

$$= s(t) + \mathrm{i}\mathcal{H}\{s(t)\} \tag{6}$$

with i being the imaginary number, $*$ the convolution operator and $\mathcal{H}\{\}$ the Hilbert transformation. The envelope $E_{\mathrm{s}}(t)$ of the data was calculated using the analytical signals $s_{\mathrm{a}}(t)$:

$$E_{\mathrm{s}}(t) = |s_{\mathrm{a}}(t)|. \tag{7}$$

Now, the maximum of the data envelope and the maximum position $t_{\max}$ at the time axis was computed as well as the FWHM of the envelope peak which contained the maximum (Fig. 2).

In a next step, a binning of the M and F data, was preformed. To determine the functions $M_0(f,T)$ and $F_0(f,T)$ for the different transducer pairs (ij) eight degree polynomials were fitted to the binned data without artificial damage:

$$M_0^{ij}(f,T) = \sum_{k=0}^{8} a_k^{ij}(f)T^k, \tag{8}$$

$$F_0^{ij}(f,T) = \sum_{k=0}^{8} b_k^{ij}(f)T^k, \tag{9}$$

with the frequency dependent polynomial coefficients $a(f)$ and $b(f)$. During binning, the sandard errors of the data combined in the binns are also determined and used for data fitting. By changing the binning width, the fit can be improved to a certain point. However, fitting without binned data is also possible, but

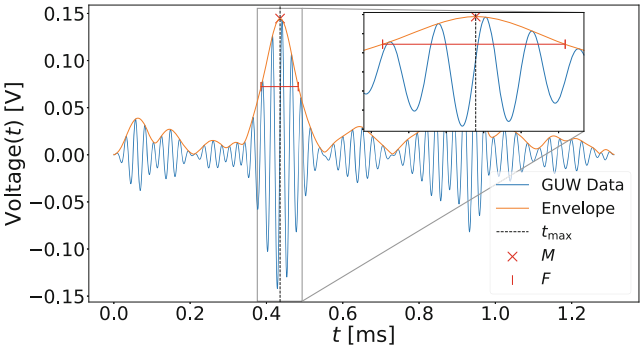


Fig. 2. Example for maximum M and FWHM F determination. The $\times$ marks the maximum of the envelope. F is computed for the envelope peak with the maximum amplitude and is illustrated as the red line. Also the location of the maximum is determined and can be seen as the dashed line in the plot. The displayed signal was recorded with transducer 1 as actuator and transducer 7 as sensor with a frequency of 40 kHz at a temperature of 21.2 °C.

this does not lead to optimal results. In the present case, the binn widths were determined by trail and error. The values of $\Delta M_0^{ij}(f,T)$ and $\Delta F_0^{ij}(f,T)$ from Eqs. (2) and (4) were computed separately for the upper and lower boundary. By computing the values of Eqs. (8) and (9) for all temperatures and adding always a value of 1e−6 to, e.g., $\Delta M_{0,\mathrm{low}}^{ij}$, if the undamaged function values of Eq. (8) are smaller than the values of the lower boundary or to $\Delta M_{0,\mathrm{up}}^{ij}$, if the undamaged function values of Eq. (8) are bigger than the values of the upper boundary. This was repeated until all data points of the undamaged case were inside the boundaries.

5 Results and Discussion

In Fig. 3a and 3b examples of data sets of M and F are shown, respectively. The M and F curves of some data sets can almost be described as linear, as can be seen in Fig. 3a for the transducer path 1 to 9. Other data sets, however, show a more complicated curve. For example, the M curve of the transducer path 1 to 10 at $f = 60$ kHz is approximately sinusoidal. The M curve of the transducer path 5 to 12 at $f = 80$ kHz is even more difficult to describe, it drops slightly from 20 °C to 25 °C and then it has a linear course up to approx. 30 °C. After 30 °C the curve takes a parabolic course until 57 °C and then it rises again slightly. In Fig. 3b, F data curves can even show discontinuities, so that it can happen that a temperature is not assigned a unique FWHM value. This can be observed, for example, for the F curve of the transducer path 1 to 9 with $f = 60$ kHz. Such discontinuities could only be observed in the FWHM data. It is assumed that this is an effect of the temperature dependence of the dispersion. Dispersion causes the modes of the GUW to diverge as they pass

through the plate. This diverging leads to a further peak detaching itself from the maximum peak of the GUW envelope, so that it can come to sudden changes in the FWHM. The temperature dependence of the dispersion leads to the modes detach at different times or locations. Because of the versatility of the data and to generalize, it was necessary to choose a feasible fit model. Since a high degree polynomial is able to represent a large part of the existing data curves, a polynomial of the eighth degree was chosen. The advantage of this model is to represent many different functions. This is made possible by adjusting $a_k^{ij}(f)$ and $b_k^{ij}(f)$ to the given data through optimization algorithms. For the purpose of proof of concept, this approach is justified even if we cannot always provide the optimal fit for the present data. To prevent data from being considered that can only be fitted insufficiently or not at all (as in the case of discontinuous functions) with a eighth degree polynomial, all fits with an R^2 value < 0.98 were excluded from the following investigations. In this case, the R^2 value was determined between the model and the binned data. To have a measure of how many fits were actually performed the fit ratio R_f of the data was determined. R_f is the ratio of the number of data sets (N_d) at which a fit trial was performed and the number of fits (N_f) actually performed $R_\mathrm{f} = \frac{N_\mathrm{d}}{N_\mathrm{f}}$. The determined R_f for $M_0^{ij}(f,T)$ can be seen in Table 2 and for $F_0^{ij}(f,T)$ in Table 3. In direct comparison between Tables 2 and 3 the $M(f,T)$ approach is much more stable than the $F(f,T)$ approach since the R_f of $F(f,T)$ is always under 0.84 and even drops to 0.08. This is due to the fact that, as mentioned above, the FWHM data curves show discontinuities. This occurs mainly at frequencies between 60 kHz and 100 kHz, but also at other frequencies the data curves often cannot be fitted.

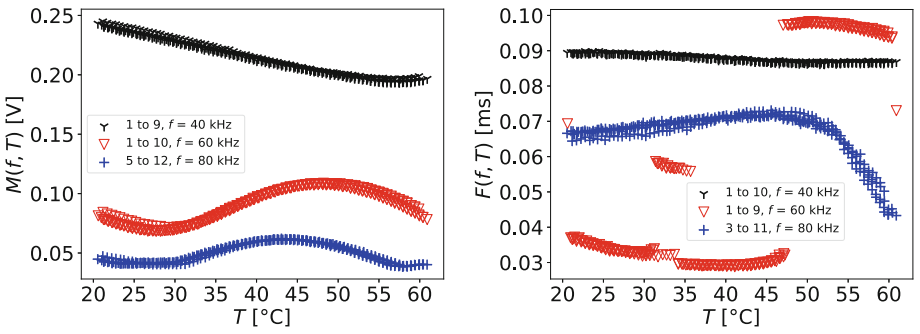
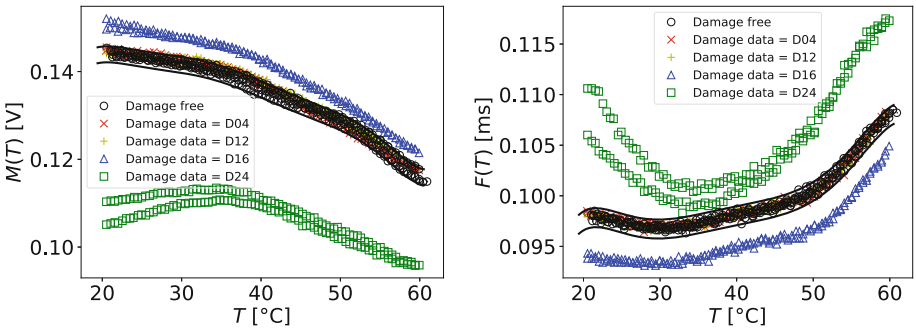


Fig. 3. Example maximum M and FWHM F data for different actuator-sensor paths (i to j) and frequencies f. Some M data have an almost linear shape while others show a more complex shape, but all data are continuous. Also F data have different shapes with continuous behavior, but there are also data with a non-continuous behavior.

In Fig. 4, the data of transducer pair 1 to 7 at $f = 40$ kHz are shown for the undamaged case (○), for damage D04 (×), D12 (+), D16 (△) and D24 (□).

Furthermore, the boundaries determined by using the method described above are also shown in black lines. Since the boundaries were determined with the data of the undamaged case, the data M and F points of the undamaged case lie within the boundaries as expected. The M and F data for damage D04 and D12 are also within the boundaries. However, in Fig. 1, D04 and D12 are far away from the transducer path 1 to 7 and was correctly identified as undamaged because there is no damage near the transducer path under consideration. This shows that even data without damage, with which the boundaries were not trained are correctly identified as undamaged. The M and F data of D16 and D24, however, are outside the boundaries over the entire temperature range. In Fig. 1, D24 is positioned directly between transducer 1 and 7 and is therefore correctly identified as damaged. Damage D12 however is further away from the transducer path and could still be detected as damage. This shows that with TS it is also possible to detect damage that is not in the direct transducer path.



(a) (b)

Fig. 4. Example maximum M (a) and FWHM F (b) data for 40 kHz and the transducer pair 1 to 7 with actuator 1 and sensor 7. With data without artificial damage (○) and data with artificial damage D04 (×), D12 (+), D16 (△) and D24 (□). The black lines represent the upper and lower boundary, within which the damage-free case is considered.

If we look at the individual measurements where $R^2 \geq 0.98$ and see how often a damage was detected correctly or incorrectly, we get the confusion matrices in Table 1. The first thing that comes to mind when looking at the Table 1 is that there are no false positive detections. This means that no damage was detected when no damage was present. This is of course due to the fact that we have determined the boundaries based on all non-damage data. This was done because of the small amount of data. The fact that the boundaries determined in this way with unknown data from a non-damaged case probably still do not detect any damage can be seen, as already mentioned, in Fig. 4. The second thing that stands out in Table 1 is that no damage was detected with a large number of data with damage. This is due to the fact that damage that is too far away

from the transducer path is not detected. Moreover, the different wavelengths of the varying frequencies are not all sensitive to the size of the damage used. In fact, there are only a few frequencies that are sensitive to the current damage, see Tables 2 and 3. With the numbers from Table 1, the true positive rate for M is 0.16 and for F a true positive rate of 0.1. This means that for M in 16% and for F in 10% of all damage cases a damage was detected. However, if we look at Tables 2 and 3, we can see that for all applied artificial damages there are still several sensor paths on which the damage was detected over the entire temperature range.

In Tables 2 and 3, the results of TS approach for all investigated data are shown. Most of the damage detections over the entire temperature range were measured with the maximum method, which is in line with expectations as this method is much more stable than the FWHM method (see number of measurements in Table 1 and R_f in Table 2 and 3), as expected since the F data also consist of measurements with non-continuous behavior. It is noteworthy that most of the detections are observed at $f = 40$ kHz. At this frequency, the wavelength is most sensitive to the 1 cm diameter damage used. Furthermore, the further away the damage is from the edge of the plate, the more damage detections can be measured over the entire temperature range. This is attributed to reflections at the edge of the plate and the resulting interferences. As described above, the signal features considered are composed of various interferences between the induced wave arriving at the sensor and its reflections at the plate edges and damage. Thus, edge reflections are part of the features, which should prevent edge reflections from being a disturbing factor in damage detection. Nevertheless, the massive reflections that occur at the plate corners seem to have an influence on the detectability, as just described. However, with TS it is possible to detect damage over the entire temperature range even in regions with strong reflections. This can be observed e.g. in Fig. 4 where both the actuator $i = 1$ and the sensor $j = 7$ are positioned directly at two plate corners. This leads to massive edge reflections at both corners, which causes interference with the induced GUW wave. Despite this, both D16 and D24 can be detected over the entire temperature range with the features M and F.

Table 1. Confusion matrices of all investigated data. Rows show if damage was present and columns show if damage was detected.

M data

Number of measurements = 410238	Damage detected	No damage detected
Damage	46358	227986
No damage	0	135894

F data

Number of measurements = 288900	Damage detected	No damage detected
Damage	19507	173693
No damage	0	95700

Table 2. Results maximum $M(f,T)$ with the R_f and the number of damage detection for the different artificial damage D04, D12, D16 and D24. Where subscript "tot" denotes total detection over the entire temperature range while subscript "par" denotes partial detection over a portion of the temperature range.

f [kHz]	R_f	$\mathrm{D04_{tot}}$	$\mathrm{D04_{par}}$	$\mathrm{D12_{tot}}$	$\mathrm{D12_{par}}$	$\mathrm{D16_{tot}}$	$\mathrm{D16_{par}}$	$\mathrm{D24_{tot}}$	$\mathrm{D24_{par}}$
40	0.97	5	22	12	19	9	26	6	27
60	1.0	3	22	4	24	1	29	2	25
80	0.92	0	17	0	20	0	20	0	15
100	0.94	0	22	0	14	0	23	0	25
120	1.0	0	23	0	28	0	28	0	28
140	1.0	0	30	0	31	0	25	0	25
160	1.0	0	27	0	25	0	18	0	21
180	1.0	0	22	0	20	0	20	0	20
200	1.0	0	26	0	23	0	24	0	19
220	1.0	0	23	0	21	0	20	0	21
240	1.0	0	27	2	25	0	29	1	26
260	1.0	2	31	5	23	0	31	2	31

To store all the polynomial coefficients a_k^{ij} and b_k^{ij} and boundary values ΔM_0^{ij} and ΔF_0^{ij} of all considered transducer pairs and frequencies a memory of 28.6 MB was needed. This amount of memory is very small compared to the memory space of 104.9 MB that just a single baseline signal of the presented data requires.

Table 3. Results FWHM $F(f,T)$, with the R_f and the number of damage detection for the different artificial damage D04, D12, D16 and D24. Where subscript "tot" denotes total detection over the entire temperature range while subscript "par" denotes partial detection over a portion of the temperature range.

f [kHz]	R_f	$\mathrm{D04_{tot}}$	$\mathrm{D04_{par}}$	$\mathrm{D12_{tot}}$	$\mathrm{D12_{par}}$	$\mathrm{D16_{tot}}$	$\mathrm{D16_{par}}$	$\mathrm{D24_{tot}}$	$\mathrm{D24_{par}}$
40	0.89	8	21	6	22	6	22	4	24
60	0.31	0	8	0	10	1	7	0	9
80	0.08	0	2	0	2	0	2	0	2
100	0.5	0	11	0	7	0	10	0	15
120	0.86	0	16	0	17	0	20	0	19
140	0.86	0	15	0	16	0	21	0	20
160	0.89	0	13	0	16	0	17	0	16
180	0.72	0	17	0	18	0	16	0	17
200	0.69	0	14	0	12	0	19	0	16
220	0.89	0	20	0	14	0	20	0	20
240	0.75	0	18	0	17	0	22	0	22
260	0.89	0	19	0	20	0	24	0	21

With TS, under certain conditions, it is also possible to perform temperature compensation without knowing the current temperature. This is achieved by replacing the temperature as a function variable by a temperature dependent signal feature $K(T)$ and thus obtaining functions $M(f, K(T))$ and $F(f, K(T))$. Where the condition holds that for each temperature value T there must be a uniquely assignable value $K(T)$, so that holds: $K(T_k) \neq K(T_l)\ \forall k \neq l$. This is the case, for $t_{\max}$ for all direct transducer paths at a core frequency of $f = 40$ kHz, see Fig. 5. It should be mentioned that there are also cases where $t_{\max}$ for certain frequencies and transducer paths has data curves which do not increase continuously but also have maxima, minima and inflection points or even discontinuities, so that in these cases $t_{\max}$ cannot be used as a substitute for T. The idea is to determine functions for the non-damage case as before, with which later measurement data can be compared. The only difference is that we do not determine the functions as a function of T but of a signal feature as a substitute. This gives us the possibility to perform the temperature compensation only on the basis of the measured GUW data. In the case of $t_{\max}$, the current measured value $t_{\max}(T)$ is passed to the functions $M(f, t_{\max}(T))$ and $F(f, t_{\max}(T))$ and their results are compared with the current measured M_m and F_m values.

As an example, the M and F data for $f = 40\,\text{kHz}$ from path 1 to 7 are shown as a function of $t_{\max}$ for the direct transducer paths in Fig. 6. In Fig. 6a, it can be seen that the D24 damage at M is again detected over the entire measurement range, similar to Fig. 4a, while the D16 damage is detected over a wide range, but from 0.45 ms onward, part of the D16 data is within the allowed boundarys. This shift in this range is probably due to the fact that damage also changes the $t_{\max}$ values, which can cause the data to shift in one direction. This happened here, so you can see that the maximum $t_{\max}$ value of D16 data is smaller than the maximum $t_{\max}$ value in case of no damage. In Fig. 6b it can be seen that for F the damage D16 is detected over the whole measurement range. Also the damage D24 could be detected over the whole range if the upper boundary would be closer to the damage free data as in Fig. 4b. This worse boundary determination is due to the fact that the non-damage data increases sharply after 0.45 ms and fans out at the same time. Thus, the non damage M values in this range determines the ΔM value and disproportionately pushes up the overall upper boundary for smaller $t_{\max}$ values. However, this problem can be solved by improved fit algorithms or by switching to residual signal calculation. With the help of all direct transducer paths, it was possible to detect all damage on the plate over the entire temperature range, despite the inaccurate boundary determination. Thus, it is generally possible with TS to compensate the temperature and to detect a damage without measuring the current temperature.

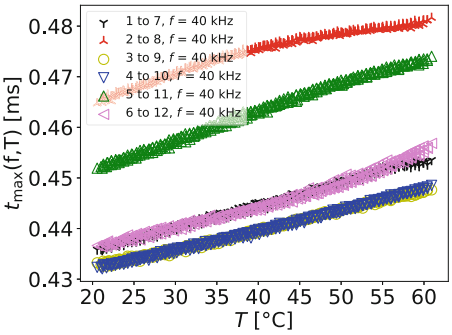


Fig. 5. In this graph $t_{\max}$ data is plotted as function of T for direct transducer paths i.e. all paths i to j for which $j = i + 6$ with $i \in \{1, 2, 3, 4, 5, 6\}$ applies. All $t_{\max}$ data for these transducer paths increase steadily with T.

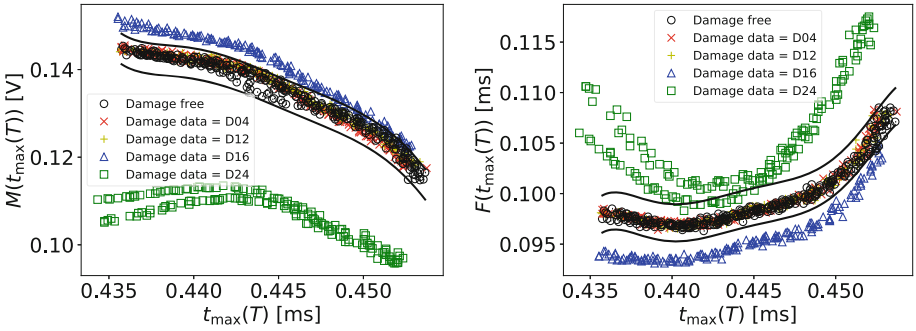


Fig. 6. Example maximum M (a) and FWHM F (b) data as function of $t_{\max}(T)$ for 40 kHz and the transducer pair 1 to 7 with actuator 1 and sensor 7. With data without artificial damage (○) and data with artificial damage D04 (×), D12 (+), D16 (△) and D24 (□). The black lines represent the upper and lower boundary, within which the damage-free case is considered.

6 Optimisation and Embedded Systems

In earlier work [13], TS was used at GUW data from [16] to determine feature vectors $\langle f, T, M, t_{\max}, F, \rangle_p$ (one for each path measurement, merged to one vector) using the unscaled signal features together with the temperature measurement. This multi-path feature vector was used as input data for an Artificial Neural Network (ANN) to predict a damage (binary classification) and its spatial position. The ANN basically approximates the introduced analytical temperature scaling (implicitly) and the mapping of the signal features on damage features (classification and localisation via the multi-path data). Although, the ANN was simple and consisted only of 24 neurons with 166 parameters, the implementa-

tion on resource-constrained material-integrated sensor nodes is a challenge. A feed-forward ANN requires the following minimal computational (γ) and memory (σ) resources:

$$\sigma^{\mathrm{ANN}} = |L_0|(3 + t_d) \cdot D + \sum_{i=1,l} |L|_i \cdot \left(2 + |L|_{i-1} + t_d\right) \cdot D \tag{10}$$

$$\gamma^{\mathrm{ANN}} = |L_0|(m_n + t_n) + \sum_{i=1,l} |L|_i \cdot \left(|L|_{i-1} \cdot (m_n + a_n) + t_n\right) \tag{11}$$

with D as the data bit width (e.g., 16), —L— as the number of neurons in a (fully connected, except input) layer i, m_n the number of machine instructions required for two-operand multiplication, a_n the number of machine instructions required for two-operand addition. The memory is measured in bits and the computational resources are expressed in the number of unity machine instructions (data transfer, addition, multiplication, division). The estimation of machine instructions required for high-level mathematical functions like the exponential function and their intermediate memory requirements is highly dependent on computer architecture and software, finally composing the transfer function requiring t_d data registers and requiring t_n machine instructions for one computation. Each neuron cell of the ANN consists of an adder unit requiring an accumulator and for each incoming edge a weight register. The output of the adder is fed into the transfer function unit.

A typical target system is shown in [19] that is suitable for material-integration. It consists of an ARM Cortex M0(+) with 8 kB RAM and 32 kB ROM storage [20]. Energy is provided via an external RFID source. Although the ARM processor core provides high performance in general, there are hard energy constraint limiting the computational time for a numerical task, and hard memory constraints. Using software-implemented floating point 32-bit arithmetic (D=32), one measuring path (one sensor node) with 5 input variables and one output variable that provides a measure for a damage detection (e.g., layer configuration is [5,8,1] with 8 hidden neurons), and using a sigmoid transfer function (td=1, computation requires only an accumulator, all other intermediate storage are stack and process registers), the ANN would require about 2944 bits or 368 Bytes memory. From the memory requirements point of view, the ANN fits in the RAM. Using the qfplib (written in ARM assembler) [21], there is $a_n \approx m_n \approx d_n$ (division) ≈ 100, $e_n \approx 200$ (exponential function), in total for the sigmoid function t_n=$2a_n$+d_n+e_n=500. The forward computation of the entire ANN requires finally about 15100 machine instructions, with a total computation time of 1ms with f_{clock}=16 MHz and $\approx$ 1 instruction/Hz. With respect to the computational time, the ANN can be processed on an embedded ARM Cortex M0 processor. Using interval-scaled integer 16bit arithmetic, the computational time is much lower.

But the ANN only approximates the earlier introduced model-based temperature-scaled feature variable computations and with a simple threshold decision algorithm the damage prediction can be achieved with much lower

memory and computational time requirements. In any case, the feature variables must be firstly computed. The computation of the unscaled feature variables require the hull signal derived commonly by the analytical signal using the Hilbert transformation ($\mathcal{H}$). The $\mathcal{H}$ can be computed using a forward and backward discrete Fast Fourier Transform (DFT/FFT). The complexity of the $\mathcal{H}$ is $O(N\log(N))$ using FFT and radix-2 FFTs require input data with a length $N{=}2^n$ and $n{=}\log_2(N)$, i.e., data padding and increased computational times are required if this condition is not met. In [13] it was already shown that the $\mathcal{H}$ is the major contribution to the overall computational time in the damage prediction system. Although, FFT is an optimised fast algorithm, one FFT computation requires approximately $5(N\log_2(N))$ arithmetic operations, with $N{=}4096$ data points this requires about 170 k instructions (or 10 ms) using 32 bit integer arithmetic, but 1s using the above floating point software emulation. The memory consumption of a FFT is about $2N\ D$ bits.

To reduce the computational time for the hull transformation, an approximation was evaluated using a rectifier (absolute function) and a low-pass (LP) filter of first order (RLP method). This approximation requires approximately $3N$ arithmetic operations and $2N$ data transfers and can be performed with integer arithmetic only without running in underflow or overflow errors. Figure 7 shows the comparison of both methods for a typical GUW signal. Firstly, it can be seen that the RLP method introduces a phase shift (as expected for a LP filter of first order). Secondly, if the RLP hull signal is phase-corrected (see Fig. 8) the error is below 5%, but still with a small ripple. The major memory requirement is the storage of the original sensor signal, i.e., ND bits. The hull computation using the RLP method can be performed in-place (in contrast, the HAS method is more challenging for in-place computation).

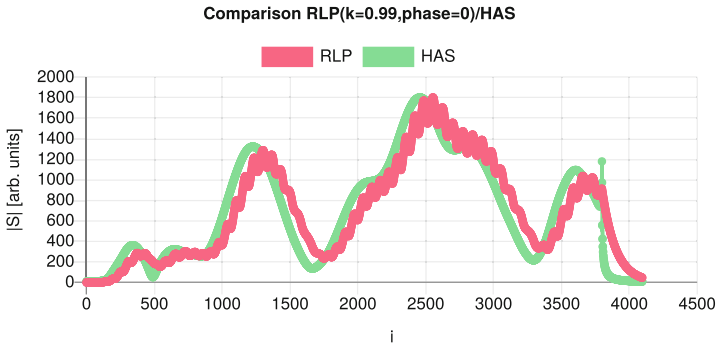


Fig. 7. Comparison of the analytical signal computed by a $\mathcal{H}$ transformation (HAS method) for a typical GUW signal with the RLP approximation method

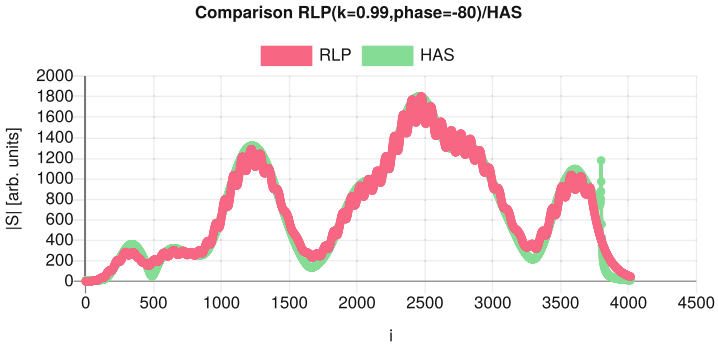


Fig. 8. Linear phase-corrected RLP approximation (shift by 80 points)

The state-based or recursive RLP function is defined as follows, transforming an input signal $s(i)$ to the hull approximation signal $S(i)$:

$$S(i) = (1 - k) \cdot s(i) + S(i - 1) \cdot k \tag{12}$$

The computation of the feature variables M is a maximum search that requires typical less than $5N$ machine instructions (one compare, two additions, two branches), and can be further reduced by defining a Region-of-Interest (ROI) candidate to narrow the time interval in which the maximum peak is expected and to provide a successive-approximation search. The computation of F requires a search around the peak and is neglectable. The temperature-scaling of the feature variables require the computation of a polynomial of degree k. The computation requires k additions and $\Sigma_{i=1,k}\ i$ multiplications.

7 Concluding Remarks

In this work, we introduced a simple but powerful method for temperature compensating in GUW data. We were able to detect damage across the entire temperature range with low computational resources. The simple polynomial model used only served as a proof of concept but could fulfill this purpose. It was also shown that under certain conditions it is possible to perform temperature compensation and damage detection with TS without knowing the actual temperature. This makes the integration of a temperature sensor in a sensor node unnecessary. The temperature-scaled features are feature variables that can be used to detect damaged in the detection range of the sensor-actuator path. Their computation using a rectifier-low-pass filter algorithm poses low computational and data storage requirements well suited for the implementation in strong resource-constrained material-integrated embedded systems. In the future, it has to be shown if with better models and fit methods the yield of used data and accuracy can be improved. If the proposed approach will prove to be practicable in real world structures depends on whether the determined feature functions

of a component can be generalised to components of the same type. Thus, the development of generalisation methods will have to be advanced in the future in order to guarantee applicability.

Acknowledgement. The authors expressly acknowledge the financial support of the research work on this article within the Research Unit 3022 "Ultrasonic Monitoring of Fibre Metal Laminates Using Integrated Sensors" (Project number: 418311604) by the German Research Foundation (Deutsche Forschungsgemeinschaft (DFG)).

References

1. Cawley, P., Alleyne, D.: The use of lamb waves for the long range inspection of large structures. Ultrasonics **34**(2), 287–290 (1996). https://doi.org/10.1016/0041-624X(96)00024-8
2. Rose, J.L.: Ultrasonic guided waves in structural health monitoring. **270**, 14–21 (2004). https://doi.org/10.4028/www.scientific.net/KEM.270-273.14
3. di Scalea, F.L., Salamone, S.: Temperature effects in ultrasonic lamb wave structural health monitoring systems. J. Acoust. Soc. Am. **124**(1), 161–174 (2008)
4. Schubert, K.J., Brauner, C., Herrmann, A.S.: Non-damage-related influences on lamb wave-based structural health monitoring of carbon fiber-reinforced plastic structures. Struct. Health Monitor. **13**(2), 158–176 (2014). https://doi.org/10.1177/1475921713513975
5. Lu, Y., Michaels, J.E.: A methodology for structural health monitoring with diffuse ultrasonic waves in the presence of temperature variations. Ultrasonics **43**(9), 717–731 (2005). https://doi.org/10.1016/j.ultras.2005.05.001
6. Harley, J.B., Moura, J.M.F.: Scale transform signal processing for optimal ultrasonic temperature compensation. IEEE Trans. Ultrason. Ferroelect. Frequency Control **59**(10), 2226–2236 (2012). https://doi.org/10.1109/TUFFC.2012.2448
7. Clarke, T., Simonetti, F., Cawley, P.: Guided wave health monitoring of complex structures by sparse array systems: influence of temperature changes on performance. J. Sound Vibrat. **329**(12), 2306–2322 (2010). https://doi.org/10.1016/j.jsv.2009.01.052
8. Douglass, A., Harley, J.: Dynamic time warping temperature compensation for guided wave structural health monitoring. IEEE Trans. Ultrason. Ferroelect. Frequency Control **1**, 3 (2018). https://doi.org/10.1109/TUFFC.2018.2813278
9. Gorgin, R., Luo, Y., Zhanjun, W.: Environmental and operational conditions effects on lamb wave based structural health monitoring systems: A review. Ultrasonics **105**, 106114 (2020)
10. Mariani, S., Heinlein, S., Cawley, P.: Location specific temperature compensation of guided wave signals in structural health monitoring. IEEE Trans. Ultrason. Ferroelect. Frequency Control **1**, 09 (2019). https://doi.org/10.1109/TUFFC.2019.2940451
11. Mariani, S., Cawley, P.: Change detection using the generalized likelihood ratio method to improve the sensitivity of guided wave structural health monitoring systems. Struct. Health Monitor. **20**(6), 3201–3226 (2021). https://doi.org/10.1177/1475921720981831
12. Mariani, S., Liu, Y., Cawley, P.: Improving sensitivity and coverage of structural health monitoring using bulk ultrasonic waves. Struct. Health Monitor. 1475921720965121. https://doi.org/10.1177/1475921720965121

13. Bosse, S., Polle, C.: Spatial damage prediction in composite materials using multipath ultrasonic monitoring, advanced signal feature selection and a combined classifying–regression artificial neural network. Eng. Proc. **10**(1) (2021).https://doi.org/10.3390/ecsa-8-11283
14. Bosse, S., Lehmhus, D., Lang, W., Busse, M.: Material-Integrated Intelligent Systems: Technology and Applications. Wiley (2018)
15. Ray, P.P.: A review on tinyml: state-of-the-art and prospects. J. King Saud Univ. Comput. Inf. Sci. (2021)
16. Moll, J., Kexel, C., Pötzsch, S., Rennoch, M., Herrmann, A.S.: Temperature affected guided wave propagation in a composite plate complementing the open guided waves platform. Sci. Data **6**, 191 (2019)
17. Paget, C.A., Grondel, S.B., Levin, K., Delebarre, C.: Damage assessment in composites by lamb waves and wavelet coefficients. Smart Mater. Struct. **12**(3), 393–402 (2003). https://doi.org/10.1088/0964-1726/12/3/310
18. Moll, J., Kexel, C., Pötzsch, S., Rennoch, M., Herrmann, A.S.: Temperature affected guided wave propagation in a composite plate complementing the open guided waves platform. Figshare. Collect. (2019b). https://doi.org/10.6084/m9.figshare.c.4488089.v1
19. Bornemann, S., Lang, W.: Experimental study on stress impact during FML manufacturing on the functional conformity of an embeddable SHM-sensor-node. Eng. Proc. **10**(1), 72 (2021)
20. Arm cortex datasheet ds10668, st microelectronics. https://www.st.com/resource/en/datasheet/stm32l031f4.pdf. Accessed 10 May 2022
21. Qfplib. https://github.com/lupyuen/qfplib. Accessed 10 May 2022

A Study on a Novel Impact Test for SHM of CFRP Based on Screwdriver Bit Launch and Electret Microphone

Cristiano Soares Junior[1], Paulo R. Aguiar[1](✉), Pedro O. C. Junior[1], Reinaldo Gotz[1], Paulo M. C. Monson[1], and Alessandro Roger Rodrigues[2]

[1] Sao Paulo State University - Unesp, Bauru, SP 17033-360, Brazil
paulo.aguiar@unesp.br
[2] University of Sao Paulo - USP, Sao Carlos, SP 13566-590, Brazil

Abstract. Carbon fiber reinforced polymer (CFRP) has recently been the subject of studies due to its increasing use in industry. Many of these studies are focused on topics related to the structural health monitoring (SHM) of this material, as its susceptibility to internal cracks and fissures are unnoticeable to visual inspection. Thus, this work presents the study of two reproducible methods of acoustic emission for two different structural conditions of a CFRP sheet in order to verify the effectiveness of each method in identifying damage, using an electret microphone for monitoring. One of the used methods is the Pencil Lead Break (PLB), which is well known in the literature, and the other is the proposed method, which consists of launching a screwdriver bit on the structure surface. To this purpose, tests were conducted for both methods on the sheet surface before and after the addition of a small metallic body to simulate structural damage. The acoustic waves generated from the tests were captured by the electret microphone attached close to the edge of the CFRP sheet. The results showed that both methods were able to identify the damage on the structure, and the proposed method is viable as a higher acoustic power alternative for monitoring the structural integrity of this type of material.

Keywords: Screwdriver bit · Electret microphone · SHM · CFRP · Monitoring

1 Introduction

CFRP has been increasingly used in the aeronautical, marine and automotive industries [1]. The main reason that makes CFRP so attractive to industry is because it is a material that has high stiffness, low weight, and high fatigue strength [2]. Although there are several advantages in using this kind of material, CFRP is susceptible to a particularly damage mechanism known as delamination. It is a critical phenomenon that introduces separation in two or more composite layers, decreasing mechanical resistance significantly [3]. Inside this scope, studies about the structural integrity of CFRP components are very valuable, since this material is increasingly inserted in structures that failures are catastrophic, such as passenger airplanes and military helicopters [1]. Research works on CFRP structural health monitoring through acoustic waves can be found in literature.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 379–386, 2023.
https://doi.org/10.1007/978-3-031-16281-7_36

The work [4], for instance, used acoustic emission signals and infrared thermography for real-time monitoring of the milling process with the purpose of estimating the surface quality and predicting defects during the process. The authors in [5] used acoustic emission waves generated from quasi-static tension tests on CFRP composites with the main objective of investigating the effectiveness of time-frequency based methodologies, such as Short-time Fourier transform, Wavelet transform and Hilbert-Huang transform, in a signal simulation analysis for damage evolution monitoring. Despite these studies, there are yet gaps to be filled about the characterization methods of this type of material due to its particularities. Thus, this work aims to study two methods of acoustic wave generation for monitoring the structural condition of a CFRP sheet using a low-cost electret condenser microphone. One of the used methods is the PLB, which is well known in the literature and used in many scientific works, such as the work conducted by [6]. The second method is the one proposed in this paper, which consists in the vertical launch of a screwdriver bit on the structure under test. The results of both methods are compared and analyzed in the time and frequency domain. What makes this work innovative is the proposal of a new method of acoustic wave generation and the acquisition of these waves by an electret condenser microphone for SHM of CFRP.

2 Material and Methods

2.1 Experimental Setup

The testing setup is illustrated in the images of Fig. 1. To perform the experiments, an electret condenser microphone model CMA-4544PF-W from CUI Devices, with 9.7 $\pm$ 0.2 mm diameter was attached close to the edge of the CFRP sheet (dimensions: 448 $\times$ 106.3 $\times$ 5.5 mm and mass of 391.46 $\pm$ 1.03 g), as shown in (2) in Fig. 1. Under the CFRP sheet a foam sheet was used, as in (7) in Fig. 1, with the same dimensions to avoid any vibration or external noise during the tests. To attach the electret microphone to the surface of the CFRP board, silicone adhesive was used around the microphone in order to not interfere in the direct contact of the transducer with the structure. The graphite break and screwdriver bit impact were performed at 80 mm from the electret microphone. The screwdriver bit used was a PZ1 type, as shown in (5) in Fig. 1, which measures 25 mm in length and weighs 4.67 $\pm$ 0.058 g. A plastic tube was used as a guide and height reference for each launch, as in (4) in Fig. 1, whose diameter is 17.45 $\pm$ 0.61 mm, height of 76.68 $\pm$ 0.02 mm, and weight of 3.17 $\pm$ 0.06 g. The acquired signals were collected and stored by an oscilloscope model DL850 from Yokogawa®, as in (8) in Fig. 1, at a frequency of 5 MHz. The electret microphone was connected to an amplifier, as shown in (1) in Fig. 1. All tests were performed at a temperature of 24 $\pm$ 2 °C.

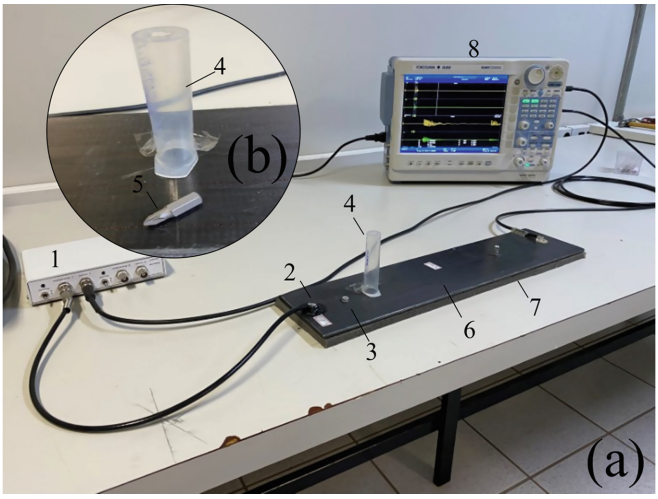


Fig. 1. (a) Testing setup: amplifier (1), electret microphone (2), metal nut (3), cylindrical plastic tube (4), screwdriver bit (5), CFRP plate (6), foam sheet (7), oscilloscope (8).

2.2 Procedures

PLB Tests. The PLB method consists of breaking a 0.5 mm thick graphite at a 45° angle to the CFRP surface. The break releases energy in the form of an elastic wave that propagates through the structure [7]. The PLB tests were conducted at 80 mm from the microphone, where initially three consecutive breaks were performed, creating the baseline (intact structure). Then, a metal nut of 7.87 ± 0.05 mm diameter and 1.178 ± 0.001 g mass was fixed 40 mm from the microphone using an ethyl cyanoacrylate-based glue, simulating damage to the plate, and then three more graphite breaks were performed.

Screwdriver Bit Impact Tests. The screwdriver bit launch test consisted of initially launching the PZ1-type bit, with the tip pointing downward, on the CFRP intact surface (baseline), at the same location as the PLB test, that is, 80 mm from the microphone, where three repetitions were conducted. The launches were carried out at an average height of 64.2 mm using a cylindrical plastic tube as a guide, as shown in (4) in Fig. 1(b). Then a metal nut of 7.87 ± 0.05 mm diameter and of 1.178 ± 0.001 g was attached with ethyl cyanoacrylate-based glue, at 40 mm from the microphone, simulating the damage, and three repetitions were performed.

2.3 Signal Processing

The digital signal processing was performed by means of MATLAB® software, and consisted initially in checking each of the test repetitions, and comparing those under the same condition in order to assure their reliability. Then, the signals were segmented in order to extract only the signal related to the test in progress. For the screwdriver

bit launch, the extracted signal refers only to the first touch of the screwdriver bit on the CFRP sheet surface, thus, removing signals caused by the subsequent rebounds. To measure the repetitions quality, after the signal extraction, the correlation coefficient was calculated for each possible pair of signals acquired for each condition of the CFRP surface. Then, the frequency spectra were extracted and compared in order to verify any variation in the frequency component of the signal between the baseline and the damaged condition. Then, the RMS values were calculated for each repetition, and the average of the three RMS values obtained for each CFRP condition was calculated. This processing was performed initially for the raw signals without any filtering, and then for the filtered raw signals in a frequency band selected for each type of test. The filter used was a Butterworth IIR bandpass, order 5.

3 Results and Discussion

Initially the signals were checked regarding the test repeatability. For this purpose, the correlation between the signals of each test was computed, as applied in the work of [8]. A correlation of 93% was obtained for the signals from the screwdriver bit launch tests and of 73% for the signals from the PLB tests. Then, the signals were properly extracted and segmented for digital processing, as mentioned in Sect. 2.3.

Figure 2 shows the raw acoustic signals acquired from the first repetition of each method for both the baseline and the damage condition. Initially, the large amplitude difference between the signals obtained from the PLB test and the screwdriver bit impact test can be observed, where the amplitude of the screwdriver bit launch is about 2.2 times larger. In both approaches, transient oscillations are observed due to the elastic acoustic waves caused by the very short duration generating sources.

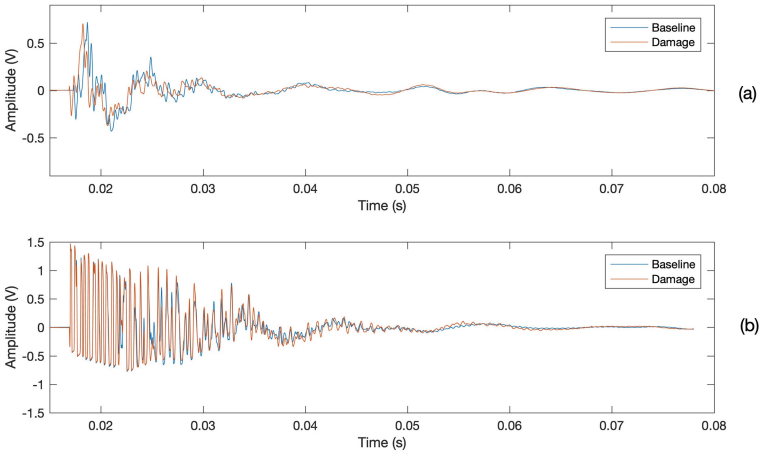


Fig. 2. Raw acoustic signals: (a) PLB; (b) Screwdriver bit impact.

On the other hand, it is observed that the signal of the bit impact has much higher oscillations, possibly due to vibrations caused in the CFRP structure by the impact,

because this method imposed much more energy in the structure, which was verified previously by much higher signal amplitudes. Finally, it is observed that there is an important difference between the signals from the baseline and from the damage condition, which opens possibilities for further analysis in both time and frequency domain for the structural health monitoring of CFRP.

Figure 3 presents the frequency spectra of the acoustic signal obtained from the PLB test. As observed in the raw acoustic signal, the amplitudes between the baseline and damage condition have important differences. However, only by searching through the signal spectra it is possible to find frequency bands in which these amplitude differences are more evident, such as the frequency band from 3 to 5 kHz detailed in Fig. 3, clearly identifying the structure condition. It is also noticed that the PLB test produced frequency response of approximately 10 kHz.

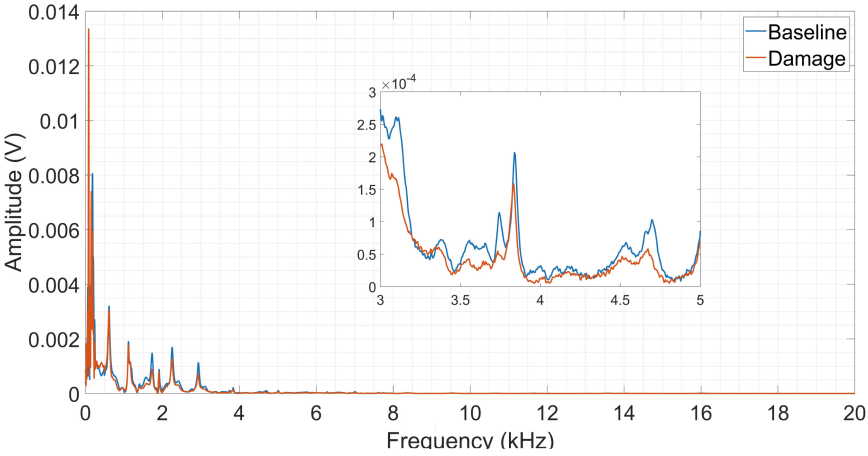


Fig. 3. PLB test frequency spectra

Figure 4 shows the frequency spectra obtained for the screwdriver bit impact test signal. It is clearly seen that the bit impact produced a much wider frequency range when compared to the PLB test, and the oscillations previously seen in the raw signal are now clearly observed in this spectrum. Thus, as in the PLB test, there is an amplitude difference in certain frequency bands, but that can only be more clearly observed when looking for them along the spectrum. The frequency band magnification from 6.5 to 7.2 kHz is shown in the same Fig. 4, in which the amplitudes differences between the baseline and the simulated damage condition are clearly observed. Thus, the bit impact test also proved sensitive to identifying the structure's condition when analyzing a selected frequency band of the spectrum.

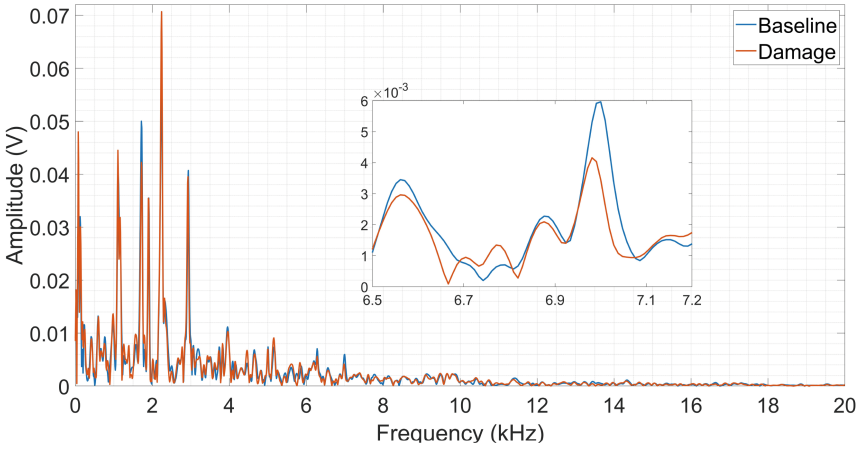


Fig. 4. Frequency spectra of screwdriver bit launch tests

The RMS mean values from raw acoustic signals were obtained for both methods. Initially, the RMS mean values were obtained along the duration of each test, as shown in Fig. 5. Amplitude differences are observed for the two structure conditions (baseline and damage) for both the PLB test and the bit impact test, being 5.93% for the PLB and 6.65% for the bit impact. However, for the PLB test the mean amplitude value is higher for the baseline compared to the damage condition, while for the bit impact test the mean amplitude value is lower for the baseline compared to the structure damage condition.

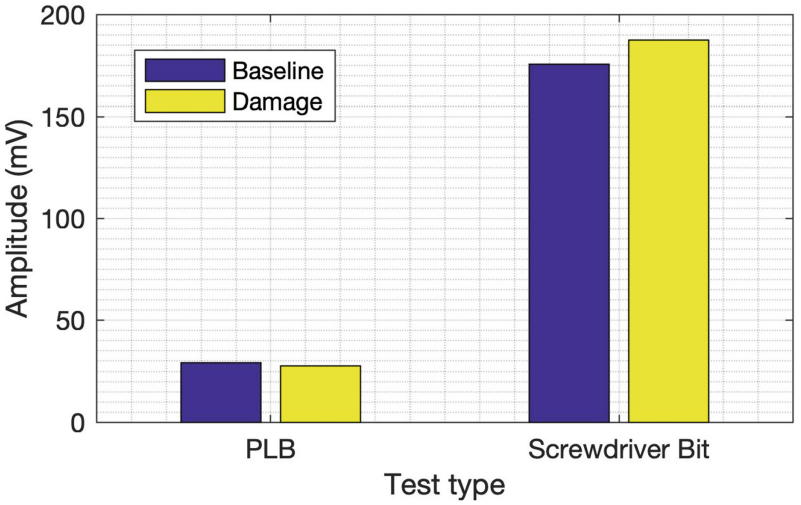


Fig. 5. RMS mean values for the PLB and Impact bit tests from the raw signal

Figure 6 shows the RMS mean values obtained for the raw filtered signal in the frequency bands selected earlier, i.e., from 3 to 5 kHz for the PLB test and from 6.5 to

7.2 kHz for the bit impact test. Similar to the previous analysis, important differences are observed between the baseline and the structure damage condition. However, now with the raw signal filtered in the selected frequency bands these differences are percentually higher, i.e., 52.2% for the PLB test and 26.1% for the bit impact test. Furthermore, it is observed that the baseline mean RMS value was higher compared to the damage condition for both tests. This is possibly caused by the attenuation of the elastic acoustic waves in these selected frequency bands when adding the metal nut to the structure.

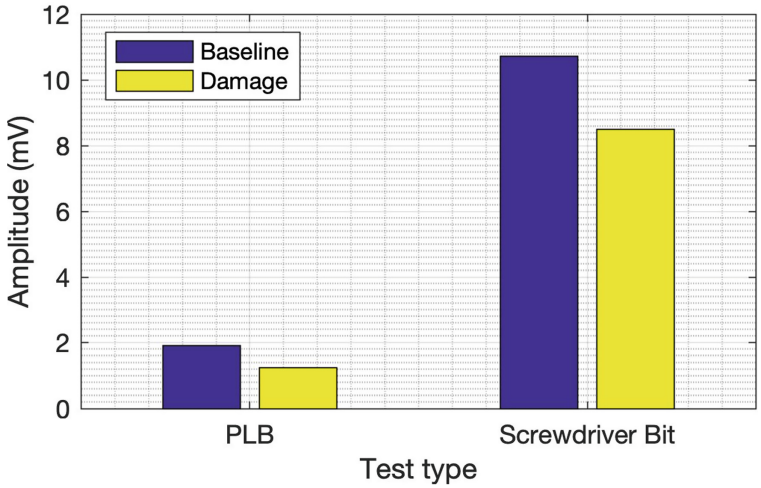


Fig. 6. RMS mean values for the PLB and Impact bit tests from the raw signal filtered in the selected frequency bands

4 Conclusion

This work proposed a new method for CFRP structural monitoring based on launching a screwdriver bit on the surface and capturing the acoustic waves by an electret condenser microphone. The PLB method was also studied for comparison. The raw signals from both methods showed good repeatability revealed by the high correlation, and the amplitude differences between the baseline and the damage condition in the structure were observed. It is also noted that the proposed method's amplitudes are up to 2.2 times higher than the PLB test, and showed more accentuated oscillations along the signal transient, possibly due to the vibration caused by the bit impact on the structure. From the frequency spectra analyses, it was found that the proposed method signal responded to a frequency range approximately twice higher than the PLB signal. However, it is necessary to look for frequency bands in the spectra in which amplitude variations are evident between the baseline and the damage condition. Thus, the 3 to 5 kHz frequency band for the PLB signal and the 6.5 to 7.2 kHz frequency band for the bit launch signal were selected to represent the structure condition. Finally, the average RMS value from the raw signals was analyzed without and with filtering in the selected bands, clearly

demonstrating that the best effectiveness of the methods is obtained by using the filtered raw signals in the selected frequency bands. It is worth mentioning that no damage or visible mark was observed due to the impact of the bit on the surface. Thus, the proposed method has shown to be a viable alternative for evaluating the condition of the CFRP structure. Further studies may consider other signal processing approaches, more detailed evaluations of signal frequencies related to structure condition, different materials, real damages, and improvements in the experimental setup, such as bit type and launch height.

References

1. Wandowski, T., Malinowski, P.H., Ostachowicz, W.M.: Delamination detection in CFRP panels using EMI method with temperature compensation. Compos. Struct. **151**, 99–107 (2016). https://doi.org/10.1016/j.compstruct.2016.02.056
2. Altin Karataş, M., Gökkaya, H.: A review on machinability of carbon fiber reinforced polymer (CFRP) and glass fiber reinforced polymer (GFRP) composite materials. Def. Technol. **14**(4), 318–326 (2018). https://doi.org/10.1016/j.dt.2018.02.001
3. Barile, C.: Innovative mechanical characterization of CFRP by using acoustic emission technique. Eng. Fract. Mech. **210**(November 2017), 414–421 (2019). https://doi.org/10.1016/j.engfracmech.2018.02.024
4. Luiz Lara Oliveira, T., Zitoune, R., Ancelotti, A.C., da Cunha, S.S.: Smart machining: monitoring of CFRP milling using AE and IR. Compos. Struct. **249**(October 2019), 112611 (2020). https://doi.org/10.1016/j.compstruct.2020.112611
5. Lu, C., Ding, P., Chen, Z.: Time-frequency analysis of acoustic emission signals generated by tension damage in CFRP. Procedia Eng. **23**, 210–215 (2011). https://doi.org/10.1016/j.proeng.2011.11.2491
6. de Almeida, V.A.D., Baptista, F.G., de Aguiar, P.R.: Piezoelectric transducers assessed by the pencil lead break for impedance-based structural health monitoring. IEEE Sens. J. **15**(2), 693–702 (2015). https://doi.org/10.1109/JSEN.2014.2352171
7. Ribeiro, D.M.S., Aguiar, P.R., Fabiano, L.F.G., D'Addona, D.M., Baptista, F.G., Bianchi, E.C.: Spectra measurements using piezoelectric diaphragms to detect burn in grinding process. IEEE Trans. Instrum. Meas. **66**(11), 3052–3063 (2017). https://doi.org/10.1109/TIM.2017.2731038
8. de Castro, B.A., de Melo Brunini, D., Baptista, F.G., Andreoli, A.L., Ulson, J.A.C.: Assessment of macro fiber composite sensors for measurement of acoustic partial discharge signals in power transformers. IEEE Sens. J. **17**(18), 6090–6099 (2017). https://doi.org/10.1109/JSEN.2017.2735858

Optimization of Non-destructive Damage Detection of Hidden Damages in Fiber Metal Laminates Using X-ray Tomography and Machine Learning Algorithms

Chirag Shah[1], Stefan Bosse[2(✉)], Carolin Zinn[1], and Axel von Hehl[1]

[1] Institute for Materials Engineering, Chair of Materials Science and Testing, University of Siegen, Paul-Bonatz-Straße 9-11, 57076 Siegen, Germany
[2] Department of Mathematics and Computer Science, University of Bremen, 28359 Bremen, Germany
sbosse@uni-bremen.de

Abstract. Detection of hidden damages in Fibre Metal Laminates (FML) is a challenge. Damage detection, classification, and localization is a part of the lower levels of Structural Health Monitoring (SHM) and is critical for damage diagnosis. SHM is an extremely useful tool for ensuring integrity and safety, detecting the evolution. Early damage detection and understanding of damage creation can avoid situations that can be catastrophic. X-ray tomography is a powerful tool for research as well as damage diagnostics. But high-resolution tomography results in high measuring and computational times up to 10 h for one specimen. The paper presents an early method of accessing the sections of the FML for identifying internal damages using X-ray imaging by optimized and adaptive zooming and scanning using automatic Region-of-Interest extraction with Machine Learning methods. The generated knowledge and the image data collected would further accelerate the development in the field of autonomous SHM of the composite and hybrid structures like fibre metal laminates which would further reduce the safety risks and total time associated with structural integrity assessment. A comprehensive image-based data set is collected by means of X-ray CT images containing micro-scale damage mechanisms (fibre breakage, metal cracks etc.) and macro-scale damages like delaminations. Starting point is an image sets were measured with two different X-ray CT devices with a static parameter set (set in advance and a-priori) and posing many limitations and issues that make damage diagnostics difficult. The adaptive and iterative measuring process should increase the quality of the images and decrease the measuring time significantly.

Keywords: Process optimization · Structural Health Monitoring (SHM) · Machine Learning (ML) · X-ray computer tomography (X-ray CT) · Fibre metal laminates (FML)

1 Introduction

Automated detection of hidden damages in laminate materials is still a challenge. Commonly, imaging techniques require the intervention and analysis by experts. X-ray CT

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 387–402, 2023.
https://doi.org/10.1007/978-3-031-16281-7_37

imaging techniques allow the inspection of hidden damages. Major limitations are image resolution, noise, contrast, intensity inhomogeneity, and X-ray diffraction and reflection effects. Beside algorithms for damage detection, measuring time is finally a major limitation of tomography. The measuring time typically increases with the image resolution. In [1], a rigorous taxonomy of damage patterns were investigated and applied to X-ray CT data for automated damage detection. Although, automated damage detection could not be validated with the proposed data processing algorithms, suitable intermediate features could be derived that can be used to define Regions-of-Interest (ROI). One reason that prevents robust and generalized automated damage detection was the low quality of the X-ray images with all limitations mentioned above. The measuring time for one sample was about one hour. In this work, we try to optimize the measuring process with respect to measuring time and image quality proposing a ROI-based adaptive measuring process that provides full CT images with coarse resolution and zoomed images of spatially bound ROIs with high resolution. The ROI prediction is performed at measuring time and provides measuring parameters at real-time.

Structural Health Monitoring has become a significantly vast area of research due to the increasing need of implementing a viable solution for non-destructive health monitoring. The SMH includes sensing technology, data acquisition, transmission and management, and health diagnostics [2]. SHM has led to a breakthrough by integrating the aspects of computer science and technology to make the structure have self-sensing and self-diagnostic abilities.

For a civil infrastructure to be in an operational status, it is very important to identify and detect the internal damages which could be detrimental to the overall performance of the structure and could potentially lead to a catastrophic failure. One step in this direction could be facilitated through the X-ray CT investigations of these structures leading to damage detection. X-ray computer tomography is becoming increasingly important among the non-destructive inspection techniques for applications where the three-dimensional (3D) nature of the phenomenon is important, or where the evolution of critical feature is of interest, either during manufacturing or under in-service conditions. Unique insights can be gained from the X-ray CT investigation revealing the damage patterns and in-service degradation. Although the capabilities of the X-ray systems have evolved over a period of time, specifically the laboratory-based equipment, its application in analyzing composites and hybrid structures still remains to be challenging [3]. This could be attributed to the thin cross-section of the composites as compared to the length or width of the specimens (non-isometric). A variety of these internal damages could be classified on the basis of certain consistent physical damage characteristics (such as type, location, size etc.) structured on different hierarchical levels with different differentiation. These damage patterns could further be utilized for GUW measurements. GUW propagation changes with the changes in the in-state characteristics of the specimens. A change in propagation pattern can be recorded with the corresponding unique damage pattern. This would allow comprehensive damage identification as these damage levels could be further correlated with the sensor response of the ultrasonic guided waves in order to enable differentiated damage identification and finally a class assignment of sensor responses leading to damage diagnosis.

A prerequisite for classification of these damages is the detectability and in addition, the identification of the specific damage characteristics. The identification of these damages is also dependent on the resolution limit. The detection and resolution limits are influenced both by the magnitude of the damage involved and the ability of the method used. Therefore, the paper highlights the findings from the X-ray CT investigations of different fibre metal laminates with artificially created defects to understand and determine the detection and resolution limits. In this paper, it is an attempt to develop the sensing technology for the monitoring of damage patterns that are unable to be detected with visual inspection or naked eyes using X-Ray CT derived 3D image volume sets. These 3D image volume sets are further utilized by the ML algorithms to identify and detect any anomalies when compared with the baseline undamaged specimens to optimize the measuring process. Figure 1 shows the schematic description of the research work highlighting the relation between detection and resolution limits, damage classification, and damage class identification using GUW signals which is a later goal of the research work.

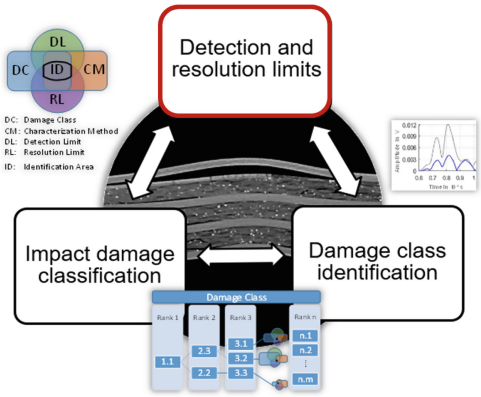


Fig. 1. Schematic description of the research work.

The research work also highlights a use-case study of automated damage detection using ML algorithms by investigating GLARE specimens with artificial defects. Glass-reinforced aluminium, known as GLARE, is a fibre metal laminate (FML) consisting of alternating S2-glass/FM94-epoxy composite plies and 2024-T3 aluminium layers. It is extensively used in the aerospace sector, where this detection technique can help to identify the damages and optimize the maintenance of aircraft. Aluminium materials under tensile loads are sensitive to the level of load and the type of variation of the load level. For this reason, crack growth rate, as well as residual strength (when the crack has developed), guides the selection of an appropriate alternative material candidate for aluminium structures [4]. GLARE offers higher intrinsic fatigue crack growth resistance [5] which has been a major driver for its selection as upper fuselage panels and skins by Airbus [4, 6]. The fatigue crack growth rates in GLARE are considerably lower in comparison to monolithic aluminium under identical loading conditions and are approximately constant for the major part of the loading [7]. GLARE has a shorter

crack initiation life but a remarkably longer crack propagation life in contrast to the monolithic aluminium where the fatigue life consists mainly of a long crack initiation phase and a small crack propagation phase. The laminated layout of FML also creates a material with good impact and damage tolerance characteristics [8]. The inner metallic layers are protected from corrosion by the fibre/epoxy layers whereas the fibre/epoxy layers are protected by the metal layers from picking up moisture [9].

2 X-Ray CT Investigations of Fibre Metal Laminates

2.1 Principle of X-ray Analysis

With the emergence of X-ray tomography for inspection of defects in parts, it became possible to inspect final parts and reject them based on defect size or location, according to some criteria. This has become routine for industrial inspection of castings, injection mouldings and composites as summarized in the review of industrial applications of X-ray tomography [10]. The non-destructive nature of the method allows the investigation of internal defects such as porosity and cracks in parts along with the internal details of the samples [11]. FML specimens with a variety of artificially created damages were fabricated to understand their detection using X-Ray CT methods. For this purpose, FML GLARE specimens were suitable for the study due to their relatively smaller density when compared to steel laminates allowing a better transmissivity of the X-rays through the specimens enabling better exploitation of the penetrating power of the high-density focused X-rays.

Feature detection (damage) using X-ray CT methods is highly dependent on the differences in contrast between the constituting elements within the specimen being investigated. For composite and hybrid materials it is related to differences between matrix, fibres, the constituting metal plies and the defects from the manufacturing (such as embedment of the foreign particle). In absorption mode X-ray CT, the contrast arises from the differences in linear attenuation coefficients (μ) of these constituting elements. In the range of X-ray energies used for the composite materials, the attenuation coefficients for a specific point (x, y, z) within the material, is given by [12, 13]:

$$\mu(x, y, z) = K\rho\frac{Z^4}{E^3} \tag{1}$$

where K is a constant, ρ is the density, Z is the material atomic number and E is the energy of the incident photons.

Consequently (directly proportional),

$$\mu(x, y, z) \propto K\rho\frac{Z^4}{E^3} \tag{2}$$

This means that materials with low atomic numbers exhibit low X-ray attenuation. This could be advantageous as X-rays can be transmitted through large composite materials but could also result in a poor contrast between fibres and matrix. It also limits the detectability of narrow matrix cracks. However, as a non-destructive tool, X-ray CT can provide 3D information to assess the quality of the manufactured components despite issues regarding the achievable spatial resolution specially in the examination of large components [3]. Figure 2 shows the schematic of X-ray computer tomography process.

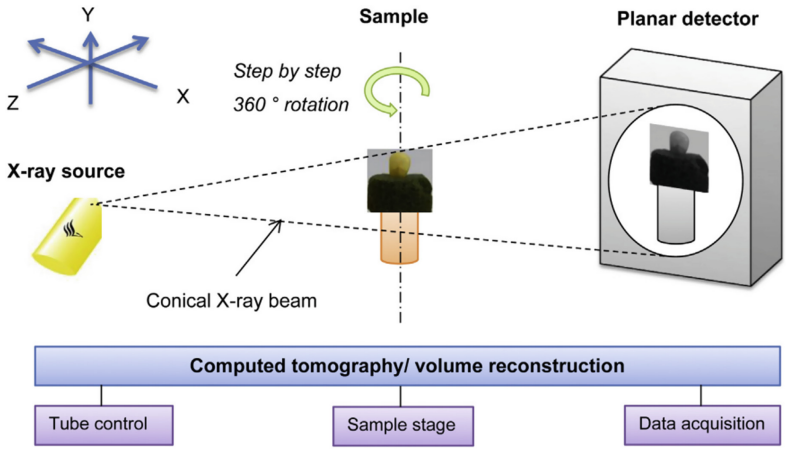


Fig. 2. Schematic of the X-ray micro-computed tomography process, taken from [14]

2.2 Experimental Procedure of X-ray CT Analysis

Detection and resolution limits of the X-ray CT method as a suitable characterization method are the prerequisite for accurate damage classification. These limits are determined by scans with various artificially induced damages. Since the characteristics and the exact location of the artificially induced damage are known inside the specimens, a manual finding of the damage and the analysis of the recorded CT images is ensured. To understand the detectability of a variety of damage patterns, various artificial damages were created to replicate the real case damage patterns occurring inside fibre metal laminates. These damages include:

- Disbonded laminate layers (i.e., delamination, kissing bonds) and weak bonds.
- Fibre breakage,
- Metal cracking and
- Sensors and sensor nodes (initially dummy sensors).

A 15 × 50 cm^2 GLARE 3-3/2 plate was fabricated consisting of 15 regions embedded with a variety of different artificial defects to replicate cracks in metal layer, fibre breakage, and delamination. Figure 3 shows the fabricated FML plate with different artificial defects. These regions with defects were cut into smaller 5 × 5 cm^2 specimens for evaluation in X-ray CT. The total thickness of the fabricated plate was 1.72 mm with a metal volume fraction of 70%.

Fig. 3. Fabricated GLARE plate (left) and specimens cut out from the plate to be investigated using X-Ray CT (centre) and a schematic figure of the layup configuration (right).

GFRP refers to Glass Fibre Reinforced Plastic prepreg consisting of FM 94-27%-S2-Glass-187-460. 0° and 90° refer to the layers of unidirectional glass epoxy oriented in a cross-ply configuration. Figure 4 shows the X, Y and Z planes for reference in an FML specimen. XY plane refers to the top and bottom face of the specimen and the Z plane refers to the thickness of the specimen. The X-ray CT investigations of these specimens were carried out using the General Electric Phoenix v|tome|x M system located at the Fibre Institute, University of Bremen. The specimens were stacked together for the CT investigations to overcome the limitations posed by the thin specimens. Thin specimens or plate-like structures, in general, are not well suited for X-ray CT investigations as the reconstruction algorithms rely heavily on the specimens being scanned to be of isometric shape.

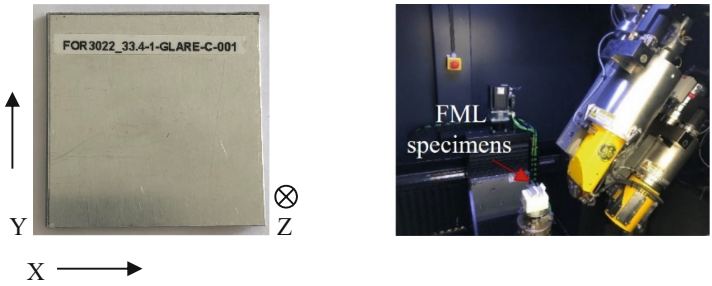


Fig. 4. A cut out FML specimen to be investigated using X-Ray CT (left) and the placement of the specimens inside the chamber for the investigations (right).

This ensures uniformity in terms of the absorption of X-rays through various sides and faces of the specimens. With plate-like structures, the penetration of X-rays varies based on the side of the plate being exposed to the X-ray source. The top face (XY plane in Fig. 4) of the specimen being scanned while facing the X-ray source would involve the scan through the thickness (X-ray penetrating through-thickness; Z-plane) which is usually quite smaller in comparison to the length and width of the specimen.

This means X-rays have a smaller path to travel to penetrate through the thickness whereas this is not the case when the specimen is being scanned through the width or the length. This creates ambiguity in terms of the X-ray intensities detected at the detector screen. Due to this, the final reconstruction of the specimen is irregular and could lead to inappropriate and irrelevant results.

2.3 Results of X-ray CT Analysis

A challenge with the X-ray CT analysis of the FMLs is associated with the energy required to fully penetrate the metal layers while obtaining the required contrast between the different density materials [15]. Defects related to cracks in the aluminium layer, foreign object embedment, resin fraction inhomogeneities and delamination were detected which are the key defects related to different failure modes encountered in FMLs. These CT investigations were carried out to assess its ability to provide information in 3D, qualitatively or quantitatively in a non-destructive fashion. The figures below illustrate the types of defects that can be imagined non-destructively by X-ray CT (Figs. 5, 6, 7, 8, 9 and 10).

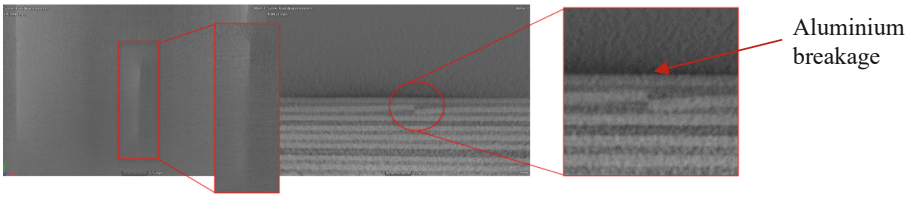


Fig. 5. X-ray CT image of the specimen with an artificially created crack in aluminium layers

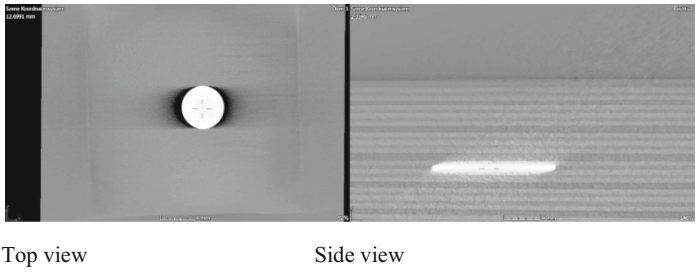


Fig. 6. X-Ray CT image of the specimen with embedded dummy sensor.

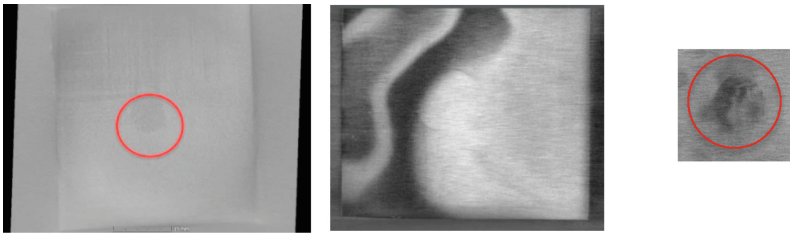
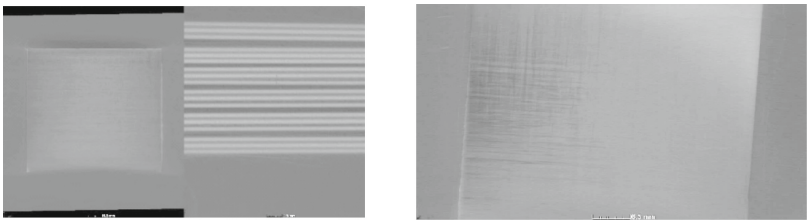


Fig. 7. X-ray CT image of the specimen with Teflon foil in a red circle (left), delamination (dark regions) due to Teflon (centre) resulting from Teflon foil and resin inhomogeneities (dark patches; in the right image) in top view [1].

The detectability of prepreg curing defects, fibre breakage, and defects resulting from liquid entrapments are challenging. Since such defects do not relate to any change in the



Top view Side view

Fig. 8. X-ray CT image of the specimen with water inclusion (left) and universal ballistol oil inclusion (right) in top view.

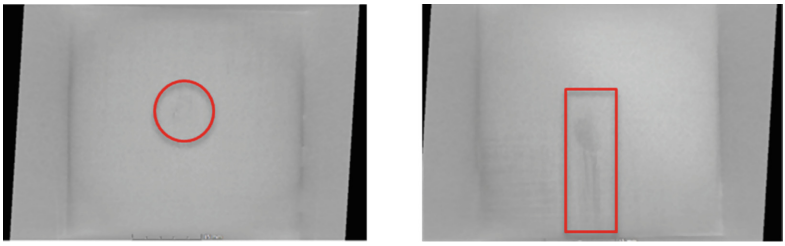
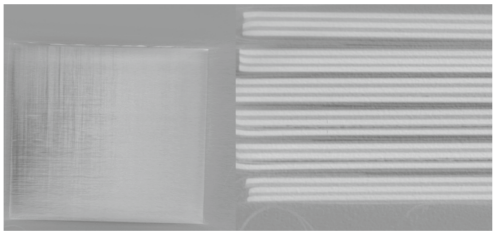


Fig. 9. X-ray CT image of the specimens with localized contact prepreg heating in top view.



Top view Side view

Fig. 10. X-ray CT image of a specimen with Loctite Frekote (mould release agent) inclusion.

density of the local region of interest, their detectability with the X-Ray CT methods is quite difficult. An automated X-ray CT image analysis using ML methods is further performed to investigate if the actual region of interest (ROI) is detected by means of non-monitored incremental learning or clustering (in Sect. 3).

3 Optimization of X-ray CT Measurements with ML Methods

The previous section gave an overview of specimens with different damage patterns and the challenges to identify them robustly in the CT images. To optimize the measuring process with respect to measuring time and image quality, we propose a ROI-based

adaptive and zooming measuring process that provides full CT images with coarse resolution and zoomed images of spatially bound ROIs with high resolution.

The X-ray CT image data quality is defined by:

1. Resolution ρ (spatially averaged) in lines/mm and sharpness ζ;
2. Homogeneity of image intensity distribution $I_{\text{avg}}(x, y, z)$ over the entire spatial volume (3D), i.e., material with same density and physical X-ray interaction properties should deliver the same intensity;
3. High contrast $C = I_{\max}/I_{\min}$ for the minimum and maximum material density;
4. Low diffraction and reflection artefacts A;
5. Low noise I_{noise}.

Beside these quality properties, the image quality is defined by a set of statistical and aggregate variables like mean, maximum/minimum, sharpness, contrast, spatial deviation, frequency spectra and MTF properties, intensity distribution, and most important texture features with specific geometric constraints (lines, points). These aggregate variables are closely related to the measuring parameters that should be optimized. In [16] the authors conclude, that "sharpness, contrast and noise were determined as a function of the number of projections. The number of projections was found to affect the contrast and the noise most, and had much less influence on resolution". The resolution of a CT imaging system can be specified by the Modulation Transfer Function (MTF) [17].

The X-ray CT measuring process is characterized by a static and a dynamic parameter set, P_{s} and P_{d}, respectively. The static parameter set defines the limitations of the imaging device itself, i.e., of the X-ray source (e.g., spot size) and the detector (sensitivity, converters, number of pixels and resolution), e.g., influencing the quality of the X-ray collimation. The dynamic parameter set includes the size if the specimen with respect to the field of view angle, the specimen rotation interval ($\alpha \in [0, 360]°$) and rotation increment $\Delta\alpha$ defining the projections, and the X-ray intensity and energy. In [1], three different ML-based algorithms were introduced and evaluated to find damage indicator features all applied to Z-profile CT image data (i.e., Z-axis slices of the image volume with respect to the specimen surface):

1. Variational Auto-encoder (VAE) delivering a 2D feature image indicating an anomaly if the reconstruction error is high and therefore providing a damage indication (used primarily for ROI selection and will be introduced in the next section);
2. Convolutional Neural Network (CNN) delivering a 2D binary damage classification image;
3. Self-organizing Kohonen maps (SOM) delivering a 2D similarity image, i.e., marking regions that pose some z-signal correlation.

A damage feature can be directly related to a damage pattern (probability) or just being an indicator for an anomaly that indicates there is maybe a damage nearby the spatial position and require further analysis either by data processing algorithms (ML) or by experts. A damage indicator correlates closely to the output of the anomaly detector.

The objective of the proposed ML-based measuring process optimization is two-folded:

1. Real-time adaptation of the dynamic measuring parameters to improve the image quality (max C, max ζ min I_{noise}, min A);
2. Reduction of measuring time and improving image resolution by performing an iterative zooming scan process by changing the rotation and translation with respect to identified ROIs, finally adapting the field of view angle and specimen position with respect to the X-ray beam.

3.1 Anomaly Detector as a Feature Marking Function

An anomalies detector is proposed to be a suitable feature marking function for damage and defect diagnosis. Input data for the proposed anomaly detector is three-dimensional CT image data of a layered material structure (here: FML plate). This anomaly detector is trained by unsupervised ML methods with baseline data (non-damaged specimen or an undamaged part of a specimen). The detector function should mark image regions where the Z-profile differs highlighting the region of interest (Anomalies Region-of-Interest, AROI), used for the following adaptive and iterative measurement process. These marked features can (but need not) identify damages or material defects. This detector is implemented with an auto-encoder architecture that reconstructs an encoded version of a signal to its original signal. Figure 11 shows the pre-processing workflow of the original raw CT image data with image transformation, image cropping, creating of a 3D Image data matrix cube, and finally Z-profiling with cylinders cutting z-signals from the image layer.

The variational auto-encoder (VAE) is implemented with a recurrent state-based Long-short Term Memory (LSTM) network. In this work, the first prototype consists of two LSTM layers with each consisting of about 7–9 LSTM neuron cells, and three dense layers with one neuron left and right from the LSTM layers.

The Z-Profile signals should contain local damage or defect information of any kind as a difference to a baseline signal. A Z-Profile signal is an average over a cylinder of

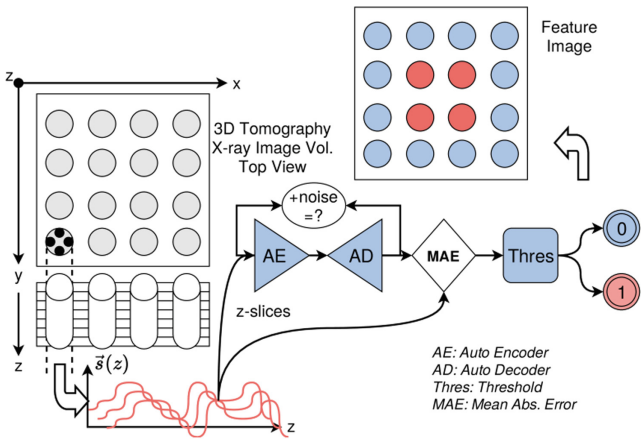


Fig. 11. Schematic representation of the slicing and Z-profiling of the 3D volumetric X-ray CT image sets and a variational auto-encoder with LSTM neural network architecture [details in [1]).

radius R. Samples of Z-signals from two images A0 (baseline) and A1 (defect; pseudo delamination) at various (X, Y) positions are shown in Fig. 12.

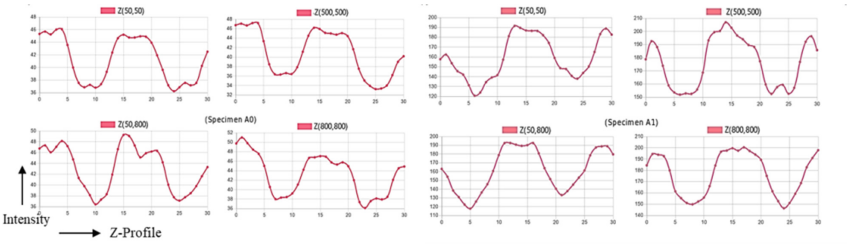


Fig. 12. Examples of z-signals of the baseline and specimen with defect at identical locations.

Some results of the anomaly detector are shown in Fig. 13. The AE was trained with all Z-signals over the circle-R-segmented (x, y) space of the image data cube of the baseline image. This trained AE was then applied to all Z-signals over the circle-R-segmented (x, y) space of the image data cube of the defect image consisting of pseudo delamination which could not be viewed manually. The delamination damage of the plate results in a large feature marking area. In [1] a rigorous evaluation showed that the VAE approach could not be used stand-alone without the correlation and comparison of the output of other methods delivering additional feature information, discussed in the next section. One major issue with the training of VAE is a spatial inhomogeneous intensity distribution and image artefacts (based on X-ray diffraction and reflection), which should be minimized by the following adaptive and optimized measuring process.

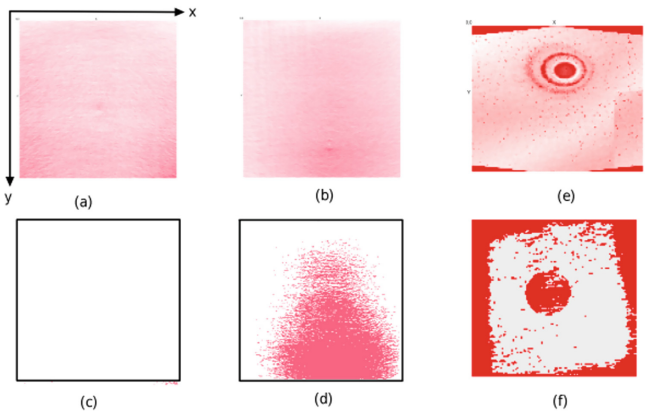


Fig. 13. Feature marking results of the anomaly detector with visible difference in image intensity scaling of both the baseline and defected specimens (a, b) with a delamination damage, results after threshold filtering (c, d), for an impact damaged specimen (e), and for a resin wash-out damaged specimen (f)

3.2 Adaptive and Iterative Measuring Process

The acquisition time for the specimens showed in the previous sections scanned at Fibre Institute in Bremen with GE Phoenix device was about 30 min. The one specimen with a higher resolution was scanned at MAPEX with a Zeiss X-radia system and had an acquisition time of around 6 h. This high measuring time is a major issue and challenge in any scanned X-ray NDT.

There are two types of magnification that can be varied: Mechanical magnification and Optical magnification. Mechanical magnification (GE Pheonix) works on the principle of the distance between the source and the object being examined. Here the magnification is controlled based on the principle of the distance between the source and the object. This is controlled manually. Optical magnification (Zeiss Xradia) is basically dependent on the use of a scintillator which converts the X-ray image into visible light which is then magnified using an optical lens.

For the reconstruction, the sample needs to be fully placed in the field of view. The number of projections has impact on the spatial resolution and the measuring time [16]. The reconstruction only considers the area which is being investigated and if the entire sample is not placed in the field of view, you will not get the complete reconstruction of the specimen being investigated. Field of view is also directly related to the resolution of the final results. Higher the field of view, lower the resolution and lower the field of view, higher is the resolution.

In general, flat surfaces and plates like structures like fibre metal laminates are not considered to be good enough for the X-ray CT because parallel surfaces are not properly penetrated by the X-rays which leads to image artefacts and lack of detail in the data set, particularly in the plane of the flat surface parallel to the beam. This can be seen in the CT results that we already have. This was the reason why we had to stack up the specimens altogether to be investigated.

The adaptive and iterative measuring process algorithm consists basically of alternated scanning and numerical processing steps, relying on a variable zoom technique proposed in [18]:

1. Perform a fast 2D transmission X-ray measurement ⇒ $\alpha=0$ with specimen surface orthogonal to the X-ray beam;
2. Perform statistical and texture analysis to optimize measuring parameters;
3. Apply damage feature marking algorithms (like semantic fully convolutional networks FCN) to the root 2D image to extract ROIs by dividing the image in small segments and applying a supervised trained anomaly predictor to the segments (CNN) or apply a FCN on pixel-level; add ROIs to a ROI data base;
4. Perform a fast 2D transmission X-ray measurement ⇒ $\alpha=90$ with specimen surface parallel to the X-ray beam;
5. Repeat step 2, add ROIs to data base;
6. Perform an iterative fast coarse-grained 3D CT scan with $\Delta\alpha=10°$ with dynamic parameter adaptation optimizing the image quality in real-time;
7. Perform statistical and texture analysis to optimize measuring parameter;
8. Apply VAE, CNN, and SOM algorithms to z-profiled CT data (see [SHA22] for details);
9. Fusion of (VAE, CNN, SOM) output to extract ROIs and add ROIs to ROI data base;
10. Perform high-resolution scanning of selected ROIs by choosing appropriate field of view angels and rotation settings (increment and interval of α).

Alg. 1. Basic iterative and zooming scan algorithm

3D Z-profiled CT image volumes can be processed by any kind of damage classificator (VAE, CNN, SOM, DT) using the z-signals as input and applying the predictor to each pixel in the x-y plane. 2D CT images can be processed by an image segmentation (using image segment as input data for CNN) or by using image transforming algorithms like FCN using the entire image as input and delivering the entire feature map image as output.

The control of the measuring process M adapting the measuring parameters is an optimization problem considering a quality function $Q(I)$ that should be maximized and the measuring time $T(M)$ that should be minimized by segmenting the entire specimen volume in smallest segment volumes given by the ROI selection process:

$$\begin{aligned} &\arg\max_{p} \; Q\left(\hat{I}(M)\right) \\ &\arg\min_{p} \; T(M) \end{aligned} \tag{3}$$

The overall measuring process and data processing chain is shown in Fig. 14. The output of the 3D CT image feature marking process is an intermediate 2D damage feature indicator image (spatially orientated parallel to the specimen surface). A point clustering algorithm (e.g., DBSCAN) is finally applied to the fusioned feature image to get relevant ROIs updating the ROI data base.

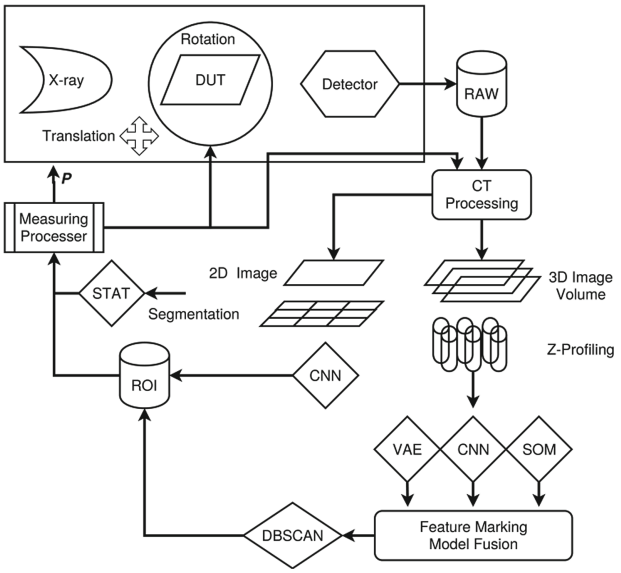


Fig. 14. Overview of the adaptive and iterative X-ray CT measuring process

Typical damages investigated in [1] were either locally bound (impact or resin washout damages) and therefore well suited for the ROI approach, or pose an extended large area (delamination) without clearly detectable geometric bounds. The measuring time can be reduced by ROI selection by at least 50% in the first case. If only one major damage pattern should be identified, delaminations can be detected fastly with low-resolution images by an anomaly detector. Therefore, an ordered set of damage patterns (with respect to geometrical features and expected size) should be analyzed iteratively with different resolutions and hence measuring times. The measuring and analysis process can be stopped if one damage pattern can be identified, modifying the above processing flow:

```
1. Perform 2D X-ray measurement (one projection);
2. Search damages that are extended in x-y and z-axis direction;
   if found go to end;
3. Perform first ROI selection; if ROis found go to 6;
4. Perform 3D X-ray measurement with low resolution of
   entire specimen;
5. Perform second ROI selection; if ROIs found go to 6;
   else go to 8;
6. Perform 3D X-ray measurement with high resolution of
   selected ROIs;
7. Perform damage analysis; if damage found go to end;
8. Perform 3D X-ray measurement with high resolution
   of entire specimen;
9. Perform global damage analysis.
```

Alg. 2. Advanced iterative scanning algorithm using an ordered set of damage patterns to be searched

4 Summary and Outlook

The detectability of certain damage patterns such as curing defects, disbonds and delamination due to liquid ingression visually using X-ray CT methods remains challenging. Image quality and measuring times of several hours are limiting factors for the deployment of X-ray CT damage diagnostics in FML materials. A ROI-based adaptive zooming iterative scanning process was proposed to improve the quality of the CT images and to reduce the measuring time significantly. The investigations of different composites and hybrid materials with a variety of damage patterns require advanced multi-level predictor models, starting with anomaly detection on the lowest level (for ROI selection), damage classification on mid-level, and damage localization on upper levels, furthermore, increasing its overall accuracy and consistency in identifying different damage patterns using selected ROI image volumes. The proposed measuring process optimization has to be investigated and evaluated rigorously with a broad range of specimens posing different damage patterns at different spatial locations.

References

1. Shah, C., Bosse, S., von Hehl, A.: Taxonomy of damage patterns in composite materials, measuring signals, and methods for automated damage diagnostics. Materials **15**(13), 4645 (2022). https://doi.org/10.3390/ma15134645
2. Li, H., Ou, J.: Structural health monitoring: from sensing technology stepping to health diagnosis. Procedia Eng. **14**, 753–760 (2011)
3. Garcea, S.C., Wang, Y., Withers, P.J.: X-ray computed tomography of polymer composites. Compos. Sci. Technol. **156**, 305–319 (2018)
4. Pora, J.: Composite materials in the Airbus A380-from history to future. In: ICCM13 Proceedings, Paper 1695 (2001)
5. Roebroeks, G.: Towards GLARE - the development of a fatigue insensitive and damage tolerant aircraft material. Ph.D. thesis, Delft University of Technology (1991)
6. Pora, J., Hinrichsen, J.: Material and technology developments for the Airbus A380. In: 22nd International SAMPE Europe Conference of the Society for the Advancement of Material and Process Engineering, La Défense, Paris, France, 27–29 March 2001
7. Alderliesten, R.C., Vlot, A.: Fatigue crack growth mechanism of GLARE. In: Proceedings of the 22nd International SAMPE Europe Conference, Paris, France, pp. 41–52 (1991)
8. Vlot, A., Gunnink, J.W.: Fibre Metal Laminates, An Introduction. Kluwer Academic Publishers, Dordrecht (2001)
9. Alderliesten, R.C.: Fatigue crack propagation and delamination growth in GLARE. Ph.D. thesis, Delft University of Technology (2005)
10. De Chiffre, L., Carmignato, S., Kruth, J.-P., Schmitt, R., Weckenmann, A.: Industrial applications of computed tomography. CIRP Ann. Manuf. Technol. **63**, 655–677 (2014). https://doi.org/10.1016/j.cirp.2014.05.011
11. du Plessis, A., Yadroitsava, I., Yadroitsev, I.: Effects of defects on mechanical properties in metal additive manufacturing: a review focusing on X-ray tomography insights. Mater. Des. **187**, 108385 (2020)
12. Buzug, T.M.: Computed Tomography – From Photon Statistics to Modern Cone-beam CT. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-39408-2
13. Hanke, R., Fuchs, T., Salamon, M., Zabler. S.: X-ray microtomography for materials characterization. In: Materials Characterization Using Nondestructive Evaluation (NDE) Methods. Woodhead Publishing Series, pp. 45–79 (2016)

14. Schoeman, L., Williams, P., du Plessis, A., Manley, M.: X-ray micro-computed tomography (μCT) for non-destructive characterisation of food microstructure. Trends Food Sci. Technol. **47**, 10–24 (2016)
15. Léonard, F., Shi, Y., Soutis, C., Withers, P.J., Pinna, C.: Impact damage characterization of fibre metal laminates by X-ray computed tomography. In: 5th Conference on Industrial Computed Tomography (iCT) 06, February 2014
16. van Daatselaar, A., van der Stelt, P., Weenen, J.: Effect of number of projections on image quality of local CT. Dentoma Illofomaxillofacial Radiol. **33**, 361–369 (2004)
17. Rueckel, J., Stockmar, M., Pfeiffer, F., Herzen, J.: Spatial resolution characterization of a X-ray microCT system. Appl. Radiat. Isot. **94**, 230–234 (2014)
18. Nikishkov, Y., Kuksenko, D., Makeev, A.: Variable zoom technique for X-ray computed tomography. NDT E Int. **116**, 102310 (2020)

Human Machine Interaction

Time Reduction in Online Programming – An Approach to Hand Guided Teaching for Small Batch Robot Machining

Marten Stepputat[1(✉)], Florian Beuss[1], Jan Sender[1], and Wilko Fluegge[2]

[1] Fraunhofer Institute for Large Structures in Production Engineering IGP, Albert-Einstein-Str. 30, 18059 Rostock, Germany
marten.stepputat@igp.fraunhofer.de

[2] Chair of Manufacturing Engineering, University of Rostock, Albert-Einstein-Str. 30, 18059 Rostock, Germany

Abstract. In contrast to the shortage of skilled workers and the resulting need to use their capabilities as effectively as possible, repetitive tasks are still performed by hand when it comes to small batches up to one-off, especially in smaller or medium sized enterprises. Reasons, for example, are the lack of programming and automation solutions with which simple tasks can be easily and in no time teached onto a robotic application. One of the fastest possibilities of online-teaching robots without programming is hand-guiding. Provided mainly for human robot collaboration most of the suitable cobots have a restricted payload. This leads to a very limited field of application regarding machining or the handling of heavy weight tools. In contrast high-load industrial robots are very common for large-series machining, but the necessary knowledge and programming effort currently prevent the use for quickly changing tasks.

This paper presents an approach, which reduces the time for online teaching of industrial robot machining. An equipped hand guidance and additional force sensors enable the user to perform the machining operation by handling the workpiece or tool as usual and records the resulting process paths and forces for teaching. Combined with a simple graphical interface for the necessary user-input the solution is empowered to repeat the recorded machining operation according to the current setup autonomously. With a target-actual comparison of the resulting forces the automatic process can be monitored. With this holistic approach the solution is suitable either for handling tasks, common machining and part-to-tool machining on external tools.

Keywords: Human robot collaboration · Online teaching · Robot machining · Hand guidance · Woodworking · Joining

1 Introduction

The use of industrial robots in the manufacturing industry is becoming more and more important, not only as a key factor for Industry 4.0 [1, 2]. Mainly pick-and-place but also assembly tasks in large-scale production are performed by them [3].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 405–414, 2023.
https://doi.org/10.1007/978-3-031-16281-7_38

In smaller manufacturing companies, whose order structure mainly consists of small series up to one-off production the use of robotics is not yet widespread. The necessary adaption of the resulting robot paths to the respective task results in high conversion times. Particularly in the woodworking sector repetitive tasks are often done completely manually, as the required changeover and reprogramming time exceeds the actual manual machining time by far. Since there is a serious shortage of skilled workers in this area, it is important to use the available personnel capacities in the most value-adding and efficient way [4, 5]. One possibility to achieve this could be to reduce the setup time for robots to such an extent that assigning repetitive tasks to them becomes economical and personnel resources are freed up.

This fact is considered by the current development trend to massively shorten the setup time of robots through novel interfaces and programming methods [6, 7]. Visual programming with predefined tasks and the programming by demonstrating or gesture and speech recognition are in current research focus as intuitive interfaces [8, 9].

Another approach for teaching, especially designed for cobots is via hand guiding. Here displacement or force-torque sensors (FTS) provide the necessary information for the resulting movements [10]. However, currently industry-ready hand guides are mainly designed for applications with cobots. Current cobot models available on the market do not meet the requirements that the use of robotics in smaller productions facilities entails, especially in terms of load capacity/size ratio.

In addition to pick-and-place tasks robots are increasingly being used for machining. Especially in woodworking there are several approaches. In contrast to hard material machining, significantly smaller process forces allow lower rigidity, larger machining spaces and increased machining options compared to CNC metal machining [11, 12]. Due to the freely configurable end effectors, both the tool can be guided over the workpiece as a classic tool-to-part process (TtP) and the workpiece can be guided in the part-to-tool process (PtT) over a stationary tool, for example a table router. This expands the application range of industrial robots especially in woodworking tremendous.

1.1 Use Cases

To illustrate the challenges of using flexible, quickly reprogrammable robots in small manufacturing companies and to evaluate the developed solution, two use cases out of the crafts sector are described below.

Use Case 1. Repetitive tasks must also be performed in wood construction when manufacturing individual pieces. For example, wooden strips needed for a custom cabinet wall are processed on saws, bench mills and/or grinders as shown in Fig. 1. Based on analysis in smaller workshops, it can be stated that in quite a few cases up to 60 of such components are machined entirely by hand. Depending on the complexity of the components, an average processing time of 30 s is needed for each work step per part, including quality control, material provision and depositing. Frequently processed components here are wooden boards, strips and blocks of various dimensions and it can be categorized as a PtT application. While the actual machining process remains the same via processing of the entire batch, the picking as well as the discarding task must be adopted according to current circumstances.

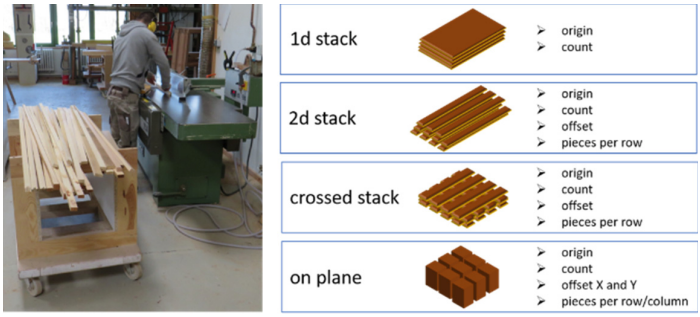


Fig. 1. Manual machining of wooden strips (left), Possible feeding and discarding strategies and necessary information (right)

Use Case 2. To also evaluate a TtP application, drilling with directions in the 3-dimensional space is considered. In this process, for example, drill holes are placed in a batch of free-form components at different points (e.g. for subsequent screw connections). Due to the complexity of the operation, by using a robot the quality and uniformity of the drilled holes can be significantly increased and the necessary human resources can be reduced. The drilling operation of one part of the batch remains the same again, like in the first use case. For processing all parts at one time within one batch, the operation must be adopted to the current position of each component.

All applications have in common that they consist of quickly changing, but repetitive tasks. The second common feature is that all processes are monitored by the worker visually or haptically. Here and especially in woodworking, the achieved result depends to a large extent on the process forces applied by the employee.

By handing these tasks over to a robotic solution, which can be taught in no time and adopted to the actual parameters and which is equipped with a suitable process control, the necessary personnel capacities could be reduced and applied to other tasks.

1.2 Current State of the Art

In the last few years, the integration of robots in production environments has been pushed further and further. Especially the development of new programming methods and interfaces has been a major topic in research and development.

In the field of machining, new approaches and solutions are stating, especially in the woodworking sector, how the additional degree of freedom and the increased movement space of the robot provide new manufacturing possibilities [11–13]. Concepts have also been developed for harder materials, such as metal, and some of these have been transferred to industrial applications [14–19]. However, the focus in metal processing is not on common milling tasks but often on finishing tasks, as the mechanical precision of the robots is usually too low for metal construction, for example in [20] and must be increased by external hardware, like in [14].

To simplify the integration effort and especially the teaching of new robot tasks, low-code or code-free applications were developed. In particular, the task-level-programming

approach allows users to define robot motion without any line of code by predefined tasks [21]. By extending the interface the parametrization of these tasks is enabled [22].

For the programming by demonstration approach the robot is usually guided via teach pendant or equivalent along the desired paths and by recording points the needed task is taught [23, 24]. Seen as a promising application for online-teaching, various concepts have been developed [6]. Especially the guiding along and recording of the required path and points has changed from moving via teach pendant to versatile solutions.

In particular, hand guidance for cobots allows moving a robot without any programming knowledge. At present the market-ready applications were almost only restricted to the manufacturers cobot series. One existing approach for heavy load collaborative hand guidance is the KEBA KeTop G10. Here the combination of 6D mouse, FTS and additional safety encoders allows to enable all benefits of cobots to heavy load industrial robots. Additionally, the implementation of the FTS allows to guide via hand-guidance or by moving the workpiece. Despite that, the solution is restricted to KEBA controlled robots. As already stated, market ready hand guiding is exclusively available for cobot series across the main robotic suppliers.

Thereof results the current research topic of the transferring hand guidance onto conventional industrial robots and enabling heavy-load HRC, like shown in [25] or [26]. A hand guiding device is already used in [21] for moving the robot while teaching.

The Ready2Pilot system from KUKA, on the other hand, also offers the possibility to teach heavy load industrial robots by hand. By using a freely positionable 6D mouse it is possible to guide any KUKA robot along a desired path. Further use, e.g. for feedback in automation solutions, is not planned yet [27].

Another approach to programming, with the focus on welding, is provided by Wandelbots with their Tracepen. Here, the necessary paths are generated by drawing them directly onto the workpiece and then were transmitted to the connected robot system. This requires additional hardware to record the paths in the 3-dimensional workspace and is currently limited to Yaskawa and Universal Robot series [28].

Most of these approaches and applications are not suitable for scenarios with very frequently changing applications in workshops, as the teaching either takes too much time compared to the actual manual processing time or requires firmly positioned hardware. The simplified task-level-programming in combination with hand guided robot motion recording offers most advantages in time-reduction and adaptability according to the presented use cases.

Nevertheless, in every solution found, the post-processing or programming effort is still too high to be used economically in workshops. So there is still a need for a solution, that makes it possible to record programmes by manual guidance without any human driven post-processing and at the same time use the existing hardware for a process control.

1.3 Objectives

From the use cases and the state of the art it can be deduced that a suitable solution must have at least the following characteristics:

- Teaching in no time with no human post- or preprocessing

- Adaptions according to current process setup without programming
- Simple process control without additional sensors
- Universal approach for applications in PtT as well as TtP or with additional tools
- Suitable as an add-on for a wide range of different existing industrial robots

This paper therefore presents a basic extension concept for existing industrial roboter that addresses these points.

The central element is a device that is used to teach processes without any programming knowledge using the programming by demonstration/hand guiding approach. By equipping it with a set of FTS the differentiation between process-forces and the resulting ones from guidance enables a simple process control. Therefore, the approach records the forces during the setup and compares them to the current measured values.

To expand the possibilities of interaction an additional HMI is also provided. It mainly consists of a touchscreen, which is located close to the teaching device to adapt the process already during recording. In order to fulfill the safety-standards of ISO 10218-1 and 2 and to cope with the circumstances in small workshops the approach contains a fence-less safety setup.

In the first part of the paper, current research trends and developed solutions are presented. Based on this in the second part the concepts for the approach are first presented and afterwards technological solutions were associated. Finally, the experimental setup and next steps are described.

2 Holistic Approach and System Layout

As shown in the previous chapter, there is no existing approach or system that meets all the requirements for quickly adaptable automation solutions for small workshops.

Therefore, a concept is presented in the following, which exactly fulfils the requirements of the use cases presented in Sect. 1.1. Based on the found user interaction and required skills the most suitable programming method or technology is derived and composed to an universal approach.

2.1 Process Analysis: Wood Machining

To demonstrate the necessary capabilities of the overall solution and in particular of the user interface, the most complex use case is used, i.e. the machining of wooden workpieces on external processing machines. It can be stated that the process with all its individual steps does not depend on the type of processing machine and that the steps in Fig. 2 are therefore generally valid. Due to the wide variety of wood types and component sizes, both the setting of the processing machine with regard to speed, necessary tools and any other parameters must be precisely adapted to the respective application and cannot be taken over by any automation solution. This also includes the post-processing, i.e. the possible dismantling or cleaning of the machining stations. This results in the following automatable steps, marked green in Fig. 2:

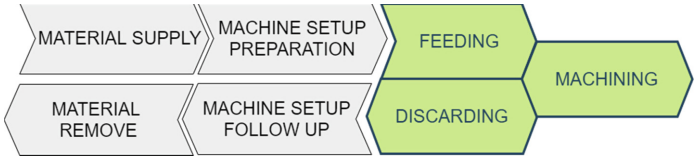


Fig. 2. Wood machining process: manual steps (gray), automatable steps (green)

In the following, necessary parameters and concepts for the respective process step will be presented.

Moving the Robot. As stated in [29] controlling a robot via a physical interaction, such as hand guiding, reduces the required knowledge and time for teaching drastically compared to common methods, like via a teach pendant. Therefore, and due to the fact, that the machining mainly contains movements along one direction in the cartesian space, a FTS based hand guidance is the central controlling element and located at the end effector. By recording the guided paths during teaching the desired movement can be played back later.

To gain a manufacturer independent approach the recording and overall communication is accomplished by an industrial PC with ROS, like stated in [30].

Feeding and Discarding. Analyses haven shown, that there is such a wide variety of raw material, that the current procedure of material supply cannot be replaced economically. As shown in Fig. 1 supply trolleys were generally used. By analyzing the possible stacking in advance the strategies shown in Fig. 1 were elaborated. To cope with the no-time-programming requirement and the fact of a restricted number of possible options the task-level-programming approach fits the most. By choosing the appropriate strategy for the current job preconfigured movements can be used. Based on simple inputs, as stated in Fig. 1, these movements can be adapted to the according circumstances. To also take into account the fact that precise material supply leads to a longer set-up time, the system is expanded with a visual object recognition. Due to the simple geometry and visible edges of the components, position deviations can be detected and the paths can be adjusted accordingly. The necessary parameters are requested via a touch screen HMI according to the currently selected strategy before picking the very first part. To avoid time consuming movements for interacting with external devices the screen is integrated in the hand guidance.

The depositing of the machined workpieces is carried out under the same conditions as the workpiece feeding.

Machining. When machining especially wooden components, the machining time and the applied force in particular have a major influence on the final result. As an example, the removal rate on a belt grinder is directly dependent on the contact pressure of the workpiece.

Taken this into account, continuous path recording and the mapping of all resulting process forces at the end effector are carried out during the teaching of the machining. By recording during the processing of the first part of the batch, the required setup time

is reduced again. Therefor the hand guidance is extended with a second FTS in such a manner, that the guidance forces can be distinguished from the resulting process forces.

These are then processed that envelope curves can be derived for automatic processing and decisions are made via a target/actual comparison, as already shown in [31]. Since in many cases the workpieces are guided in one plane when machining on external machines, the movement can be restricted to this plane via the HMI.

Safety. In the concept presented, a distinction is made between two phases: the teach phase and the automated phase.

While teaching, the entire application complies with DIN 10218-1 by having an enabling switch and an emergency stop at the hand guidance, which are directly coupled to the robot controller. Additionally the operating mode "Manual with reduced-speed" is used during the teach phase. [32].

During the automated process, both the safety towards the machine tools and the general machine safety according to DIN EN ISO 10218-1 and 10218-2 must be guaranteed. For the last part, the overall system is equipped with two laser scanners. Likewise, the use of laser scanners allows zoning so that machining can be carried out at reduced speed if a critical zone is not penetrated. This increases the availability of the system even in cramped workshops.

2.2 Additional Components and System Layout

To expand the application possibilities the system is extended with additional components as shown in Fig. 3. To interact with a wide range of possible tools and end effectors the included PLC can be controlled via the HMI. Regarding safety requirements of different types of robots all critical components are monitored by a safety PLC as well as the control of enabling and emergency stops of the applied robotic system.

In the current setup a ROBOTIQ FT 300 is used as the input FTS 1. The recorded values are processed according to Fig. 4 onto an IPC within a ROS environment. To prevent movements resulting from bias or touching the interface an exponential smoothing and a threshold is implemented.

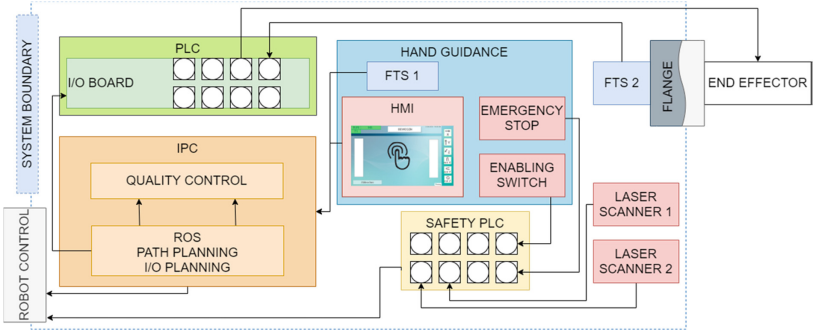


Fig. 3. Targeted system layout and necessary components

The results were processed with a factorial vector, which is also used for disabling single axes by setting the corresponding value to zero. By limiting the result to 250 mm/s and an additional exponential smoothing the final velocity vector matches to the safety regulations and a predictable movement.

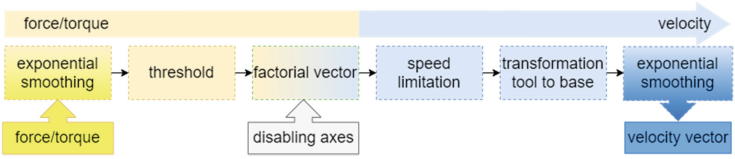


Fig. 4. Hand guiding: translation of force/torque input from the FT 300 to vectorial velocity.

For the mandatory recording of waypoints, paths and I/O interaction a time-series database onto the IPC is used. Here the current approach divides between discrete waypoints and paths. To record a path several equidistant points were stored as well as the current velocity between these points. By mapping the resulting force of the second FTS these values can be restored and used for the process control during the automation phases. The recorded movements were blended in limits according to the unknown environment and resulting risks of collisions.

3 Outlook and Conclusion

To evaluate the described approach and to provide a detailed statement about the benefit an experimental environment is currently set up. Afterwards there were several surveys planned to measure the potential and impact of the solution. A KUKA KR20 R1810 and a Fanuc CR 35iA are used for the purpose of evaluation. In cooperation with a small joinery, the designed solution is first tested for suitability using the exemplary tasks. To critically scrutinize process control, material and processing errors are forced during automated machining. The second part of the evaluation will then show the potential in time saving by comparison of purely manual machining with the automated procedure.

Theoretical Evaluation. To provide a first estimation about the benefit of the system, the first use-case is considered: Here a batch of 60 pieces of a generic workpieces is processed. Based on time measurements it can be stated that a minimal processing time of 30 s per piece is required, so that the overall machining duration is about 30 min.

By combining several time recordings in teaching with HG in other applications it can be stated that the online-approach by manufacturing the very first part will not take more than 5 min. As a result more than 25 min of personnel resources can be freed up. These capacities can be used for more complex and carpentry related tasks.

Conclusion. In this paper an approach was presented with which a quickly teachable automation solution for woodworking can be implemented. For this purpose, the necessary paths are recorded via a manual guidance for industrial robots and can then be traversed automatically. Necessary additional inputs via a HMI were pointed out and

a FTS concept enables the teaching of the robot via hand guidance on one hand and process control during machining on the other.

Regarding the distribution of processing stations in smaller workshops, a mobile platform for location-independent use is under consideration and after the proof for the tasks listed here, further possible uses of the solution are to be evaluated. Conceivable scenarios would be machining on saws or welding applications.

References

1. Javaid, M., Haleem, A., Singh, R.P., Suman, R.: Substantial capabilities of robotics in enhancing Industry 4.0 implementation. Cogn. Robot. **1**, 58 (2021)
2. Wang, B., Tao, F., Fang, X., Liu, C., et al.: Smart manufacturing and intelligent manufacturing: a comparative review. Engineering **7**, 738 (2021)
3. International Federation of Robotics IFR: World Robotics 2020
4. Brunello, G., Wruuck, P.: Skill shortages and skill mismatch in europe: a review of the literature, 12346 (2019)
5. Bundesagentur für Arbeit: Fachkräfteengpassanalyse 2019, Nürnberg, Germany (2020)
6. Heimann, O., Guhl, J.: Industrial robot programming methods: a scoping review. In: 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), p. 696. IEEE (2020)
7. Dean-Leon, E., Ramirez-Amaro, K., Bergner, F., Dianov, I., et al.: Integration of robotic technologies for rapidly deployable robots. IEEE Trans. Ind. Inform. **14**, 1691 (2018)
8. Berg, J., Lu, S.: Review of Interfaces for Industrial Human-Robot Interaction. Current Robot. Rep. **1**(2), 27–34 (2020). https://doi.org/10.1007/s43154-020-00005-6
9. Bravo, F.A., Gonzalez, A.M., Gonzalez, E.: A review of intuitive robot programming environments for educational purposes. In: 2017 IEEE 3rd Colombian Conference on Automatic Control (CCAC): Conference Proceedings, Cartagena, Colombia, 18–20 October 2017, p. 1. IEEE, Piscataway (2017)
10. Mariscal Saldana, M.Á., González-Pérez, J., Khalid, A., Gutiérrez Llorente, J.M., et al.: Risks management and cobots. Identifying critical variables. In: Proceedings of the 29th European Safety and Reliability Conference (ESREL), p. 1834. Research Publishing Services, Singapore (2019)
11. Ayari, O., Bouali, A., Méausoone, P.-J.: Cutting forces and accuracy characterization during wood machining with serial robots. Eur. J. Wood Wood Prod. **78**(4), 767–775 (2020). https://doi.org/10.1007/s00107-020-01539-4
12. Koch, J.: Der eiserne Kollege: Tischlermeister Axel Eigenstetter arbeitet mit einem Fünf-Achs-Industrieroboter von Kuka. https://www.bm-online.de/praxis-und-kollegentipps/zu-gast-beim-kollegen/der-eiserne-kollege/. Accessed 9 Dec 2020
13. Stepputat, M., Beuss, F., Pfletscher, U., Sender, J., et al.: Automated one-off production in woodworking by Part-to-Tool. Procedia CIRP **104**, 307 (2021)
14. Dryba, S., Meißner, J., Wanner, M.-C., Wurst, O.: Hochpräzises Bearbeiten von Schiffspropellern: high-precision machining of very large ship propellers. Procedia CIRP **107**, 182 (2017)
15. Brunete, A., Gambao, E., Koskinen, J., Heikkilä, T., et al.: Hard material small-batch industrial machining robot. Robot. Comput. Integr. Manuf. **54**, 185 (2018)
16. Leal-Muñoz, E., Diez, E., Marquez, J., Vizan, A.: Feasibility of machining using low payload robots. Procedia Manuf. **41**, 594 (2019)

17. Rileys Surface World: INTEC - ABB IRB 2400 ROBOTIC FINISHING CELL. https://www.rileysurfaceworld.co.uk/live/machines2/25800.pdf. Accessed 9 Dec 2020
18. Song, Y., Chen, Y.H.: Feature-based robot machining for rapid prototyping. Proc. Inst. Mech. Eng. Part B J. Eng. Manuf. **213**, 451 (1999)
19. Xiong, G., Ding, Y., Zhu, L.: Stiffness-based pose optimization of an industrial robot for five-axis milling. Robot. Comput. Integr. Manuf. **55**, 19 (2019)
20. Ferreras-Higuero, E., Leal-Muñoz, E., García de Jalón, J., Chacón, E., et al.: Robot-process precision modelling for the improvement of productivity in flexible manufacturing cells. Robot. Comput. Integr. Manuf. **65**, 101966 (2020)
21. Brunete, A., Mateo, C., Gambao, E., Hernando, M., et al.: User-friendly task level programming based on an online walk-through teaching approach. Ind. Robot Int. J. **43**, 153 (2016)
22. Steinmetz, F., Wollschlager, A., Weitschat, R.: RAZER—a HRI for visual task-level programming and intuitive skill parameterization. IEEE Robot. Autom. Lett. **3**, 1362 (2018)
23. Biggs, G., MacDonald, B.: A survey of robot programming systems, in Proceedings of the Australasian conference on robotics and automation, p. 27 (2003)
24. Billard, A., Calinon, S., Dillmann, R., Schaal, S.: Robot programming by demonstration. In: Siciliano, B., Khatib, O. (eds.) Springer Handbook of Robotics, pp. 1371–1394. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-30301-5_60
25. Fujii, M., Murakami, H., Sonehara, M.: Study on application of a human-robot collaborative system using hand-guiding in a production line. IHI Eng. Rev. **49**, 24 (2016)
26. Gopinath, V., Ore, F., Johansen, K.: Safe assembly cell layout through risk assessment – an application with hand guided industrial robot. Procedia CIRP **63**, 430 (2017)
27. KUKA Ready2Pilot. https://www.kuka.com/en-de/products/robot-systems/ready2_use/kuka-ready2_pilot. Accessed 16 Dec 2021
28. Homepage Wandelbots. https://wandelbots.com/en/roboterprogrammierung/. Accessed 16 Dec 2021
29. Gustavsson, P., Holm, M., Syberfeldt, A., Wang, L.: Human-robot collaboration – towards new metrics for selection of communication technologies. Procedia CIRP. **72**, 123 (2018)
30. Gregor, R., Babinec, A., Duchoň, F., Dobiš, M.: Hand guiding a virtual robot using a force sensor. Acta Mechanica et Automatica **15**, 177 (2021)
31. Schmatz, F., Neumann, S., Sender, J., Flügge, W., Meschut, G.: Qualitätsoptimierung im mechanischen Fügen durch Einsatz von Mensch-Roboter-Kollaboration, Stuttgart (2020)
32. DIN EN ISO 10218-1:2012-01, Robots and robotic devices - Safety requirements for industrial robots - Part 1: Robots (ISO 10218-1:2011); German version EN ISO 10218-1:2011. Beuth Verlag GmbH, Berlin

Ontology-Based Documentation of Quality Assurance Measures Using the Example of a Visual Inspection

Tatyana Sheveleva[1](✉), Kevin Herrmann[2], Max Leo Wawer[2], Christoph Kahra[3], Florian Nürnberger[3], Oliver Koepler[1], Iryna Mozgova[2], Roland Lachmayer[2], and Sören Auer[1]

[1] Leibniz Information Centre for Science and Technology (TIB), Welfengarten 1B, 30167 Hanover, Germany
tatyana.sheveleva@tib.eu

[2] Institute of Product Development, Leibniz University Hannover, An der Universität 1, 30823 Garbsen, Germany

[3] Institut für Werkstoffkunde (Materials Science), Leibniz Universität Hannover, An der Universität 2, 30823 Garbsen, Germany

Abstract. The development of a novel manufacturing process chain is a complex scientific challenge and requires interdisciplinary and inter-institutional collaboration. Data need to be exchanged continuously between involved researchers in order to coordinate between individual process steps and to identify cause-effect relationships within the process. This publication describes an approach to provide seamless digital access to quality-related data and to further structure, semantically annotate and link process- and quality-relevant data. It uses a domain-specific ontology called Visual Inspection Ontology embedded in a Knowledge Management System to support the documentation of a quality-determining process. The ontology is applied to a use case from the development of a novel process chain to manufacture multi-material shafts within the Collaborative Research Centre (CRC) 1153. A workflow to establish quality control measures regarding a novel process chain for multi-material high-performance components under development based on the proposed ontology is presented.

Keywords: Human machine interaction · Semantic annotation · Knowledge management system · Sample monitoring · Tailored Forming

1 Introduction

The ability to produce high-performance components at low-costs is a major factor when estimating the economic performance of a manufacturing company. The knowledge about the applied production processes is considered the cornerstone to ensure constantly high product quality. The use of new technologies is supporting this cause by enabling a constant acquisition of process data. Data mining methods complement the acquisition by implementing a real time analysis of the data and subsequently implementing a control system for the process. In general, this process control demands three requirements:

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 415–424, 2023.
https://doi.org/10.1007/978-3-031-16281-7_39

1. The process needs to meet the specified customer requirements [1];
2. The process parameter window resulting in the production of high quality products needs to be known; and
3. The process needs to be controlled; I.e. the process behaviour is predictable.

Process chains using known manufacturing technologies usually meet these requirements. However, they are not or only partially met within research projects developing novel production processes and technologies. The technologies used in individual process steps under development are not yet sufficiently identified, matured, and stable to guarantee the consistent quality of the manufactured components. Furthermore, the interdependencies between several process steps within a process chain are usually also unknown. The incomplete knowledge and understanding of each process step results in the inability to predict the effect of a change in a prior step on further processes. Therefore, the development of a novel process chain is considered an iterative process, adapting each production step several times and examining the effect of changes throughout the whole chain. To this mean, the collaboration of experts from different disciplines is required. The successful development of a process chain is therefore depending on an optimal data flow between the individual processes [2, 3].

The design of novel manufacturing process chains using Tailored Forming Technology is the subject of the Collaborative Research Centre (CRC) 1153, where various production methods are being investigated in various subprojects. Each subproject generates a large amount of data [4], which, when examined together, forms the basis for a cross-process quality management of a process chain. This publication presents an approach for supporting quality-assured measures within an internal Knowledge Management System (KMS) using ontology. As a use case serves a visual inspection performed before and after a manufacturing process.

2 Quality Management: State of the Art

There are several papers dealing with defects and damages of components. DIN EN ISO 9000 is used as a base for the classification of product characteristics, which are directly connected to the fulfillment of quality related requirements [5]. Additionally, Bauerdick et al. have implemented a typology of defects associated with machining processes and have described their typical geometrical and physical characteristics [6]. The development of analogy workpieces to emulate those defects is then based on the typology of the defects. Furthermore, various concepts to ensure constant quality in the production of goods and the provision of services are applied, usually based on a closed-loop control system in the form of a quality management system [7]. For these concepts, a numerical quality metric is not necessarily required, e.g. as shown for a technical complaint system [8]. The quality control can be extended from the use in subprocess tasks to complex process chains [9].

3 Background

This section gives a short introduction to the concept of Tailored Forming and presents a novel manufacturing process chain specific to this technology. Using one of the process

steps as an example, the current challenges of introducing quality assurance measures within the CRC 1153 are outlined.

3.1 Tailored-Forming and Its Specific Manufacturing Process Chain

In CRC 1153, novel Tailored Forming-specific process chains are being investigated [10]. Hereby, several processes are connected in series with the aim of manufacturing hybrid components with local-adopted properties by using different materials [11]. Due to a high number of necessary process steps, the complexity of the corresponding process chain is increased [12], which results in further challenges with regard to the reliability of the whole manufacturing process. Here, the manufacturing process chain of a hybrid shaft is presented as an example (see Fig. 1):

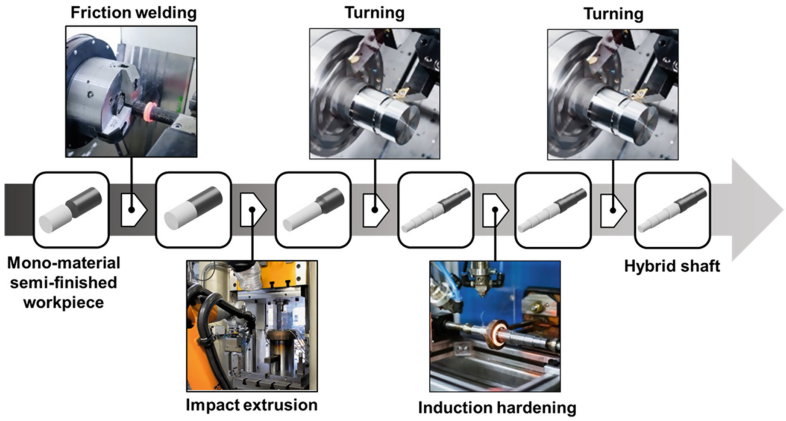


Fig. 1. Process chain for the manufacturing of a multi-material shaft by Tailored Forming

In the first step, two mono-material workpieces are joined sequentially by friction welding to a hybrid semi-finished workpiece. In the following impact extrusion, the workpiece is inductively heated and preformed to the hybrid shaft, which is then subjected to turning and, subsequently, to induction hardening. The last step is the turning of the shaft preform to the final geometry.

Within such a process chain, each process step and its parameters are determined by a responsible sub-project. In order to coordinate the process steps collaboratively and to develop the process chain as a whole, data must be exchanged bidirectionally, since the individual process steps interact with each other and a level of quality of the upstream step is necessary to perform the subsequent steps successfully. This fact underlines the importance of quality assurance measures and makes their establishment indispensable within the development of novel process chains. While quality control is usually used in established methods and seems to be a stretch for production research, there is a threat of deficiencies and pitfalls without such a system in place.

3.2 Status Quo of Quality Assurance Measures in the Given Use Case

Within the CRC 1153, various quality assurance measures are currently being used and further introduced individually by each subproject. These consider structured data, e.g. temperature, and geometry measurement as well as unstructured data, e.g. documentation of the processes, like visual inspection.

Visual inspection is performed twice during the transfer of the workpiece between the subprojects: once as output visual inspection from the giving subproject, focusing on workpiece characteristics for the own process and is optionally supported by structured as well as quantitative measurements. A non-standardized rating of the workpiece quality is usually documented in a hard copy control sheet attached to this workpiece when it is handed over to the next project. Furthermore, the workpiece is visually inspected by the receiving projects and evaluated by its suitability for its own process. The inspection by giving and receiving projects may cover different features, and the inspection results may or may not be transferred to the other subproject in a suitable data format.

Within the example process step of heat treatment, a visual inspection is performed to inspect the form and the physical appearance of the workpiece. During the input visual inspection, only visible defects are identified, e.g., obvious form distortions, pores, or cracks. It is only used to check the essential geometrical compatibility of the workpiece to the tool environment and involves identifying intolerable form deviations, e.g. length or diameter differences. Any non-visible defects like joining errors are not detectable. The output inspection is applied both with the focus on the workpiece and on the process after the heat treatment. It involves a further visual inspection of the heat-treated workpiece to identify defects induced by heat treatment, e.g. an additional form distortion caused by different thermal expansion coefficients of the used materials. The analysis of the parameters employed in case of detecting faulty workpieces is a commonly used tool to identify critical parameters and relevant remedial actions. For the heat treatment use case, a workpiece distortion is stated as a reason for the failure of the inductive hardening. Exceptional results of the inspection are noted as unstructured text on the control sheet on paper. Along the process chain, such a control sheet documents all process steps applied on the workpiece and by whom. Additionally, process-relevant information including a minimum description of the applied parameters is stored in a global sample registry, an Excel sheet on an FTP server. In some cases, data on workpieces are also forwarded by e-mail. In summary, several workflows are used to exchange data on the quality of workpieces between the subprojects. This way of working and the use of different media to pass on data about research objects are prone to errors and potentially incomplete. Not all subprojects have continuous access to quality-related data concerning their processes and the involved workpieces. This hampers the identification of cause-effect relationships within the process chain. Moreover, quality-related data is provided in different formats and in unstructured forms, which makes it impossible to search for complex data required for the use case. Based on such challenges, the following requirements have been identified in order to optimize the quality assurance measures making use of semantic annotation of the inspection-related data by the domain-specific ontology within the KMS:

1. VIOR1: The visual inspection shall be protocoled by using one medium providing the quality-related data in a structured form;
2. VIOR2: The protocoled visual inspection shall be accessible to all participating stakeholders;
3. VIOR3: The input and output visual inspections shall be protocoled separately;
4. VIOR4: The protocoled visual inspection shall provide the required data about associated samples, their defects, and persons responsible.

4 Ontology-Based Optimization of Quality Assurance Measures

The following section outlines the approach for the formal structuring of quality-related data using ontology and how it is applied within the internal knowledge management system to optimize the quality assurance measures of the heat treatment.

4.1 Design of the Visual Inspection Ontology

The potential enhancements described in Sect. 3.2 are considered within the development of the structural basis for the visual inspection protocol, resulting in the so-called Visual Inspection Ontology (VIO). Additionally, more specific requirements have been identified in joint discussions with researchers performing the inspection. These requirements focus on retrieving specific information items that the VIO is supposed to provide and that result in the following competency questions:

- Which samples have been the subjects of the input/output visual inspection?
- What types of defects have been detected within the input/output visual inspection?
- Which samples examined the input/output visual inspection in the given time period have corrosion/offset/crack/fused area?

VIO have been created using the ontology editor Protegé (see Fig. 2). It imports case-relevant classes and properties from three different existing ontologies of a high abstraction level. So, the core super-classes that are related to an observation are adopted from the Semantic Sensor Network Ontology (SSN) [13]. Due to their generalized meaning, the SSN entities form the basic skeleton of the VIO and build super-classes of the VIO-specific entities. Moreover, SSN entities serve as linking elements in the alignment of VIO with the further ontology of the highest abstraction level, the Basic Formal Ontology (BFO) [14]. The BFO enables to connect VIO to further BFO-aligned ontologies if VIO is to be applied in other research areas. A seamless alignment of the SSN to the BFO is ensured by the integration of the appropriate classes from the Provenance Ontology (PROV-O) [15]. So, the bfo:Continuant has a subclass bfo:Process, that is equivalent to prov:Activity. The bfo:Continuant has a subclass bfo:IndependentContinuant, which is equivalent to prov:Entity. The prov:Entity is a super-class of sosa:ObservableProperty, sosa:FuatureOfInterest, and sosa:Sensor. The prov:Agent is also a super-class of sosa:Sensor, and prov:Activity is a super-class of sosa:Observation. The vio:Visual Inspection, defined as a subclass of sosa:Observation, stands for the quality-related inspection performed by prov:Person. Representation of the

inspection result is realized via the usage of sosa:resultTime and sosa:hasSimpleResult. The class vio:Visual Inspection is expanded by two disjoint subclasses vio:Input Visual Inspection and vio:Output Visual Inspection, which stand for the two inspections performed before and after the heat treatment. The vio:Sample, defined as a subclass of sosa:FutureOfInterest, represents a semi-finished workpiece undergoing inspection. The vio:Defect represents the sample property, and is therefore defined as a subclass of sosa:ObservableProperty. Since the inspection is limited to detect defects without using any tools, only visible defects are considered. According to DIN 17022-2, the typical representatives of visible defects associated with heat treatment are corrosion, crack, fused area, and offset [16], which are included as subclasses of vio:Defect. However, this defect classification should not be considered complete as of now. Due to the novelty of the manufacturing technology with its high variety of the applied material combinations, workpiece geometry, and process settings, further types of defects may also occur. Once available, these have to be included into VIO subsequently to ensure a complete quality-related knowledge acquisition. A defect is characterized by its location on the workpiece (vio:isLocatedInArea). Since location is not required to be determined quantitatively by specifying coordinates, the data type xsd:string is used. It allows the relevant data to be entered as free text without any limitation in content and value. The vio:hasDefectStatus enables in turn to characterize the defect with respect to its significance for the quality of the heat treatment, like "critical" or "non-critical" (Tables 1 and 2).

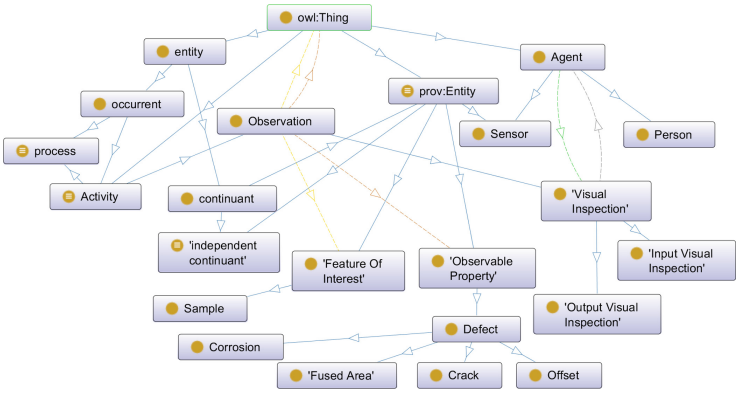


Fig. 2. Graphic representation of Visual Inspection Ontology by using Protegé's visualization plugin OntoGraf

The developed VIO is available as a Turtle file at the GitHub repository TIB Onto via https://github.com/tibonto/vio. The structure of VIO serves as a knowledge base for the creation of protocols used for documentation of inspection results in the internal KMS of the CRC1153 and thereby contributes to the establishment of a centralized quality management.

Table 1. Data type properties graphically not shown.

Data type property	Domain	Range
vio:hasDefectStatus	vio:Defect	xsd:string
vio:isLocatedInArea	vio:Defect	xsd:string
sosa:resultTime	sosa:Observation	xsd:dateTime
sosa:hasSimpleResult	sosa:Observation	xsd:string

Table 2. Identification of object properties by color.

Object Property	**Arc Color**
has subclass*	
sosa:hasFeatureOfInterest	
sosa:observedProperty	
vio:made	
vio:madeBy	
*The property **has subclass** is created by OntoGraf automatically for visualization purposes and serves as the inverse of property **rdfs:subClassOf**	

4.2 Creation of Visual Inspection Protocols in the Knowledge Management System

The internal KMS is based on Semantic MediaWiki (SMW) [17, 18], an open-source extension of the MediaWiki software [19]. It enables to collect, structure, and semantically annotate unstructured data by means of domain-specific vocabulary and thereby makes the data machine interpretable and searchable. VIO is imported into SMW by using the extension ontology2smw [20]. The key VIO entities are used in the second step to build the structure of the protocol to be presented as an SMW page. To realize this, a protocol template is created, whereby the VIO classes and properties are used as a knowledge resource. Here, the class vio:Visual Inspection is the subject of the protocol template. Other classes and properties serve in turn as its descriptive elements. Within the template, VIO classes and properties are grouped in such a form, that they are visualized as separate sections representing specific content on the Wiki page (see Fig. 3). The first protocol section captures general administrative information, e.g. inspection type (input or output), the responsible person, inspection date, the workpiece to be inspected, and inspection result. The second section is designated via the identification number of the initial inspection performed and the subject of the section content (observable properties standing for defects). It combines information about the identified defects and their characteristics (defect type, location, and status). This section is built on a query that calls the defect-relevant data after it is entered by the researcher into the system and displays it in table form. Despite the sectional splitting of the inspection-relevant data on the SMW page, both content elements are semantically interlinked with each other by being based on a common structure of VIO.

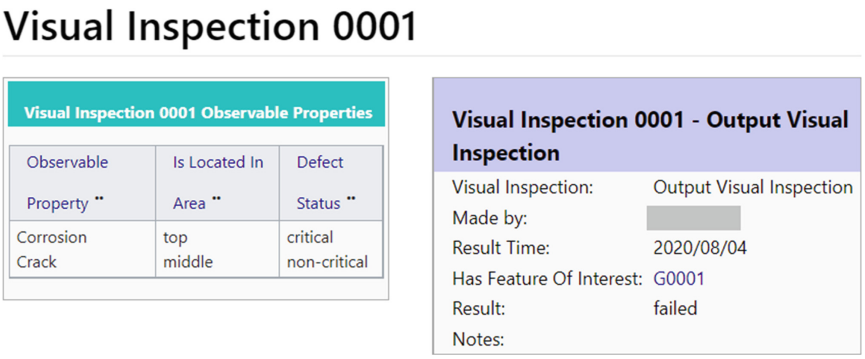


Fig. 3. Visual inspection protocol implemented in Semantic MediaWiki.

As a final outcome, using such semantic annotation of the inspection-related data, the sample control sheets on paper can be completely replaced by the digital protocols available for all subprojects along the whole process chain.

4.3 Benefits Obtained from the Use of the Visual Inspection Ontology

The benefits of the usage of the VIO can be demonstrated by the translation of the following competency question into a query that is run using semantic search: "Which samples examined within the output visual inspection in the time period between 01.03.2020 and 18.11.2020 show corrosion and whether the corrosion is critical or non-critical? Additionally, the material combination of the relevant samples is to be displayed." Applying this query, the following results are received (see Fig. 4). The query result is a table displaying a complete list of semi-finished workpieces in which corrosion has been identified after the heat treatment procedure. This information can be further used by the researchers, e.g. to identify the reason for the failure of the heat treatment procedure, possibly caused by specific behaviour of the materials. Thereby, the blue marked column entries are provided as links redirecting to the relevant Wiki page. The black marked entries represent simple values, which do not have their own pages.

The search results presented above show only one possible use case for retrieving quality-related data, that is semantically annotated using VIO. Since the internal KMS

	Output Visual Inspection	Result Time	Sample	Material Combination	Defect	Defect Status
Visual Inspection 0001	Output Visual Inspection	4 August 2020	G0001	100Cr6/AW-6082	Corrosion Crack	critical non-critical
Visual Inspection 0002	Output Visual Inspection	11 March 2020	G0002	100Cr6/AW-6082	Corrosion	critical
Visual Inspection 0003	Output Visual Inspection	18 November 2020	G0006	100Cr6/AW-6082	Corrosion	non-critical
Visual Inspection 0012	Output Visual Inspection	5 August 2020	G0226	20MnCr5/AW-6082	Corrosion	non-critical
Visual Inspection 0016	Output Visual Inspection	5 August 2020	G0228	20MnCr5/AW-6082	Corrosion	non-critical
Visual Inspection 0017	Output Visual Inspection	6 August 2020	G0229	20MnCr5/AW-6082	Corrosion	non-critical
Visual Inspection 0018	Output Visual Inspection	5 August 2020	G0230	20MnCr5/AW-6082	Corrosion	non-critical
Visual Inspection 0021	Output Visual Inspection	6 August 2020	G0233	20MnCr5/AW-6082	Corrosion	non-critical
Visual Inspection 0027	Output Visual Inspection	6 August 2020	G0219	41Cr4/C22.8	Corrosion	non-critical

Fig. 4. Results of quality-related data retrieval obtained by using competency question.

is accessible for all CRC members, the functionality of protocol documentation and data retrieval are open as well, so that depending on the use case, any other queries can be performed.

5 Conclusion and Outreach

The presented approach provides elements of comprehensive quality assurance measures within the development of novel manufacturing process chains in a large collaborative project by means of a use-case from the CRC 1153. First, all steps of the quality assurance measures have been digitized and embedded into a central KMS based on SMW, substituting former hard copy-based control sheets. By this means, visibility and accessibility of quality-related data has been significantly increased. The digital mapping of control sheets has also improved the previously rather unstructured collection of data for quality assessment. The structured and standardized collection of statements on the quality of the semi-finished workpieces has been further improved by the implementation of the VIO into the internal KMS. This enables a centralized documentation of the visual inspection procedures and their results as one quality assurance measure. Semantically annotation and linking of the data from control sheets, heat treatment parameters, and workpiece characteristics enable in a next step the analysis of cause-effect relationships between parameters within processes and the workpiece quality. The overarching availability of quality-determining process parameters broadens the perspective. Instead of optimizing a single process step and maximizing the quality of its outcome compared to the investment of resources, the perspective is planned to be extended to the whole process chain and the interaction of the sum of all process parameters identifying the most economic and highest-quality processing. As a next step, a quantitative evaluation of the influence of heat treatment parameters such as soaking times, temperatures and cooling rates onto mechanical properties like hardness and tensile strength shall be carried out regarding the quality management. An extension of the VIO covering other quality inspection methods is ongoing work.

Acknowledgements. The results presented were obtained within the Collaborative Research Center 1153 “Process chain to produce hybrid high performance components by Tailored Forming”, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 252662854, in the subprojects A2, C2, C5 and INF.

References

1. Brüggemann, H., Bremer, P.: Grundlagen Qualitätsmanagement - Von den Werkzeugen über Methoden zum TQM, 3rd edn. Springer, Wiesbaden (2020). https://doi.org/10.1007/978-3-658-28780-1
2. Altun, O., et al.: Integration eines digitalen Maschinenparks in ein Forschungsdatenmanagementsystem. In: Proceedings of the 32nd Symposium Design for X (DFX 2021), 32 (23) (2021)
3. Sheveleva, T., et al.: Development of a domain-specific ontology to support research data management for the tailored forming technology. Procedia Manuf. **52**, 107–112 (2020)

4. Mozgova, I., et al.: Research data management system for a large collaborative project. In: Proceedings of NordDesign, Lyngby, Denmark (2020)
5. DIN EN ISO 9000:2015: Qulitätsmanagementsysteme - Grundlagen und Begriffe (2015)
6. Bauerdick, Ch.: Charakterisierung und Modellierung von Bauteilfehlern – Qualitätsüberwachung, Darmstadt (2018)
7. DIN EN ISO 9001:2015-11: Qualitätsmanagementsysteme - Anforderungen (2015)
8. Linder, A., Schmitt, S., Schmitt, R.: Technical complaint management from a quality perspective. Total Qual. Manag. Bus. Excell. **25**, 865–875 (2014)
9. Goch, G., Dijkman, M.: Holonic quality control strategy for the process chain of bearing rings. CIRP Ann. Manuf. Technol. **58**, 433–436 (2009)
10. Uhe, J., Behrens, B.A.: Manufacturing of hybrid solid components by tailored forming. In: Wulfsberg, J.P., Hintze, W., Behrens, B.A. (eds.) Production at the Leading Edge of Technology. Springer, Heidelberg (2019). https://doi.org/10.1007/978-3-662-60417-5_20
11. Brockmöller, T., et al.: Rechnergestützte Entwicklungsumgebung zur Konstruktion von Tailored-Forming-Bauteilen. In: DS 98: Proceedings of the 30th Symposium Design for X (DFX 2019), Jesteburg, pp. 195–206 (2019)
12. Herrmann, K., et al.: Fertigungsstufen-basiertes Gestaltungsmodell für mechanische Bauteile. In: Proceedings of the Stuttgarter Symposium für Produktentwicklung, SSP 2021, pp. 247–258 (2021). https://doi.org/10.18419/opus-11478
13. W3C: Semantic Sensor Network Ontology (SSN). https://www.w3.org/TR/2017/REC-vocab-ssn-20171019/. Accessed 23 Mar 2022
14. Arp, R.: Building Ontologies with Basic Formal Ontology. MIT Press, Cambridge (2015)
15. W3C: PROV-O. https://www.w3.org/TR/prov-o/. Accessed 29 Nov 2020
16. DIN 17022-2:1986-06: Wärmebehandlung von Eisenwerkstoffen; Verfahren der Wärmebehandlung; Härten und Anlassen von Werkzeugen (1986)
17. Krötzsch, M., Vrandečić, D., Völkel, M.: Semantic MediaWiki. In: Cruz, I., et al. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 935–942. Springer, Heidelberg (2006). https://doi.org/10.1007/11926078_68
18. Semantic-mediawiki.org contributors: Semantic MediaWiki. https://www.semantic-mediawiki.org/w/index.php?title=Semantic_MediaWiki&oldid=76616. Accessed 24 Aug 2021
19. MediaWiki. https://www.mediawiki.org/w/index.php?title=MediaWiki&oldid=3878227. Accessed 29 Mar 20222
20. Castro, A., Strömert, P.: Ontology2smw. https://github.com/andrecastro0o/ontology2smw. Accessed 29 Mar 20222

Controlling Decisions by Head Electrical Signals

Enrico Zero(✉), Alessandro Bozzi, Simone Graffione, and Roberto Sacile

Department of Informatics, Bioengineering, Robotics and Systems Engineering (DIBRIS), University of Genoa, Genoa, Italy
enrico.zero@dibris.unige.it

Abstract. This work considers the problem of using the electroencephalogram in a real context to control devices. The proposed work takes the data from the central and parietal brain areas to perform a steady state visually evoked potentials (SSVEP)-based Brain Computer Interface (BCI) model. The BCI output was retrieved by the human head electrical activity within a scenario that requires participants to think about a predefined image. The simulations were performed by 7 healthy participants 5 men and 2 women between 23 and 56 years old. A system composed by a Neural Network has been applied to develop the predictive model. The model developed can predict the human thinking with an accuracy more than 70% in the validation set for each participant.

Keywords: Brain Computer Interface · Artificial Neural Network · SSVEP · EEG

1 Survey Literature

The advent of the Industry 5.0 has put the human behavior at the center of the industrial process [1]. With respect to Industry 4.0, the new industrial paradigm combines artificial intelligence and man common's life in order to enhance efficiency and productivity in a human-centered vision of the process [2]. Workers' security will have an even more crucial role during this development by its integration within the smart factory and the activation of systems designed for individual protection. This topic includes technologies that allow the remote intervention with the aid of virtual or augmented reality and artificial intelligence.

The interaction between human brain and external devices, namely Brain Computer Interface (BCI), started decades ago with the aim of acquiring and analyzing brain signals, usually retrieved by an electroencephalogram (EEG), in order to translate them into outputs that replace or enhance human functions [3].

Nowadays, most BCI applications focus on medical aspects, such as augmented reality [4], post-stroke motor rehabilitation [5], and even neurorehabilitation [6].

Nevertheless, many other applications are arising nowadays about non-medical uses [7,8] that exploit the potentialities of this technology. Even mobile robots [9] and cars can be controlled by the usage of BCI outputs.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 425–436, 2023.
https://doi.org/10.1007/978-3-031-16281-7_40

In this direction, the investigation of brain actions during imaginary movement represents an interesting field of research. The act of simulating a body movement requires deep thinking and uses the same principles as the ones studied in this paper. This happens because actual and imaginary movement activate similar brain areas, thus suggesting the usage of BCI for imaginary movements classification, in order to boost recovery of damaged or injured body parts. Moreover, imaginary motion is important for learning motor skills and experiments on the motion of the wrist has been analyzed [10], as well as classification of foot movements [11].

BCIs have the unique ability of tackling complex brain dynamics and extract valuable data for the classification, especially when paired with Artificial Neural Networks (ANN). They are usually implemented to manipulate the big amount of information produced by these experiments. It has been demonstrated that they can improve classification performances [12].

According to literature results, in this paper, a Time Delay Neural Network (TDNN) joint with a Pattern Recognition Neural Network (PRNN) has been used. In literature, the former has been used to identify driver's arm movements in virtual scenario [13], while the latter revealed to be helpful in recognizing speech emotions [14]. In the same direction, the monitoring of fear during driving sessions has been performed making usage of an EEG [15]. Another study focused on the detection of pilot's reactions to unexpected events [16]. The study proposed comparisons in machine learning algorithms such as Support Vector Machines, Linear Discriminant and ANN, showing which of them obtained better results.

2 Introduction

This work focuses on creating a prediction model based on the analysis of the BCI outputs retrieved by the human thought within a scenario that requires participants to think about a predefined image. The methods used for the creation of the following model is based on to the steady state visually evoked potentials (SSVEP). The SSVEP is a periodic neural response located in the subject's central visual field that is induced by a repetitive visual stimulus [17]. Several BCI applications use the SSVEP to control external systems. An hybrid BCI system that combines motor imagery-based bio-signals and SSVEPs to control the speed and direction of a real wheelchair synchronously has been set up [18]. The SSVEP was also used to develop a system which control and manage the movement of electric mobility scooters [19]. The proposed methods follow a common standard nowadays about the use of the imaginary movement to control some external devices just thinking to an image or a specific object.

The rest of the manuscript is organized in three sections. Section 3 introduces the methods and material used for this work. Section 4 reports the results obtained in the simulation. Section 5 concludes the paper and provides insights for future works about this topic. Figure 1 shows the flowchart pipeline of the paper.

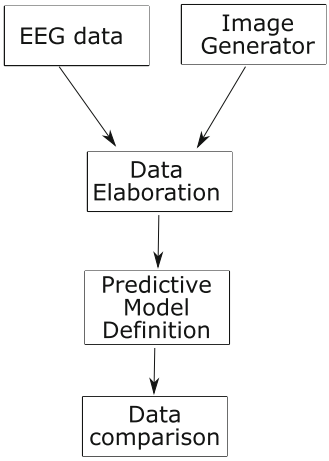


Fig. 1. Flowchart pipeline

The first step is related to the input definition and acquisition: EEG data and Image creator. All the data will be send to a Server PC for the preliminary data elaboration and the definition of a predictive model. The last step was the comparison and the performance analysis of the model set up.

3 Material and Methods

3.1 System Design

The proposed EEG-based system is composed by two sub-systems. The first one is the EEG cap system useful for the EEG signal acquisition in real time. The second one is a Matlab script which generate randomly images to be shown to each participant during the simulation. Both sub-systems send the data for the elaboration and the creation of a predictive model able to predict the image from the EEG signal to a Server PC. In Fig. 2 the system architecture is shown:

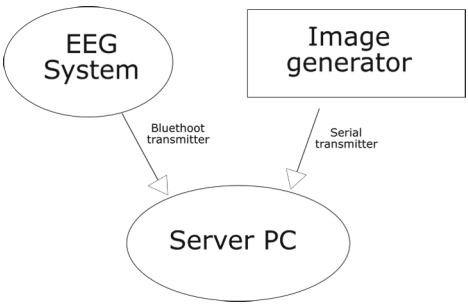


Fig. 2. System architecture

3.2 EEG Cap

The EEG Cap used in this work is the Unicorn Hybrid Black of g.tec medical engineering company. The cap has 8 electrodes located according to the International Standard System 10/20 and a sample rate 250 Hz. Besides, 2 additional reference electrodes positioned behind the ears have been considered to perform a differential filtering and to limit artifacts due to the electrical muscles' activity. The data acquisition process has been performed in Matlab/Simulink R2021b using the Unicorn Simulink Interface Hybrid Black.

In this work only 4 electrodes (F_z,C_z,C_3,C_4) located in the central and parietal regions are taken into account, because these brain areas are related to thinking and imagining movement [20]. Figure 3 shows the electrodes location during the simulations:

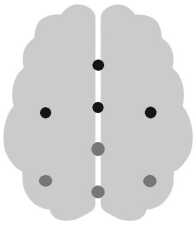


Fig. 3. Electrodes location: in black color the electrodes used during the experiment (Color figure online)

3.3 Image Generator

The image creator script is a Matlab script which show randomly two figures for 5 s, alternated with a rest image, a gray square, used to break the simulation at each trial. The images chosen must represent two contrasting scenarios (water/flame, heaven/floor ...etc.) to create opposite situations. The two images are codified with label 1 and 2.

The script also saves the epoch start and stop timestamp of each figure to perform offline a good synchronization between the two subsystems. At the end of the simulation, all these data are saved in a storage for the elaboration. Figure 4 shows the design of the environment for the image creation:

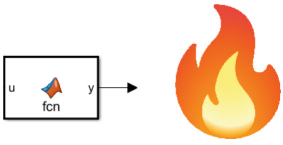


Fig. 4. Design of the environment for the image generator

3.4 Experimental Setup

The simulation was performed by five healthy men and two healthy women between 23 and 56 years old. During the experiment 5 min long, each participant must think of the image showed without blinking his/her eyes for 5 s. Between two images there are 3 s of pause where a gray figure is showed.

The participant sits in a comfortable position without moving in front of a PC screen where the image is projected during the experiment. The two figures represent a water drop and a flame.

The figures are randomly showed during the simulation. Figure 5 shows a schema of one of the 36 trials collected for participant. For each participant, two simulation has been acquired in the same day after 5 min: the first one for train and test the predictive model and the second one to validate it.

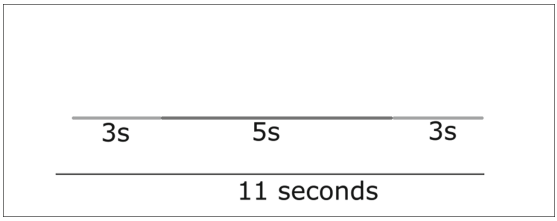


Fig. 5. Pattern of the duration of a trial

An example of a simulation is shown in Fig. 6:

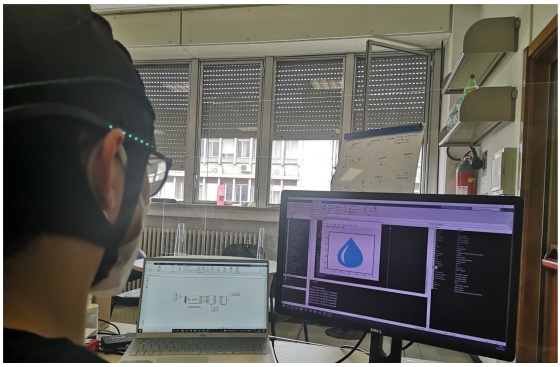


Fig. 6. Simulation environment

3.5 Data Elaboration

A pre-processing elaboration has been performed to synchronize and clean the EEG signal collected from artifacts and noise. To reduce the noise given by the electrical frequency of devices [21] a bandstop/notch filter between 49 and 51 Hz has been applied. In according with [18] a band-pass filtered between 0.1–30 Hz has been carried out. All the previous elaborations have been performed for the four channel chosen only. To validate the BCI model a supervised learning method has been implemented. The data are divided in two parts: training and test set, respectively 70% and 30% of the total. The model used for the prediction is an ANN as implemented in [22]. The learning of ANNs relates to the method of modifying the weights of the connections among the nodes of a specified network. The type of ANN defined in this work comprises: input layers where the inputs of the problem are received, hidden layers where the relationships between inputs and outputs are defined and setted by the weights and output layers where the outputs of the problem are defined [23]. In Fig. 7 is showed a generic graph representation of an ANN.

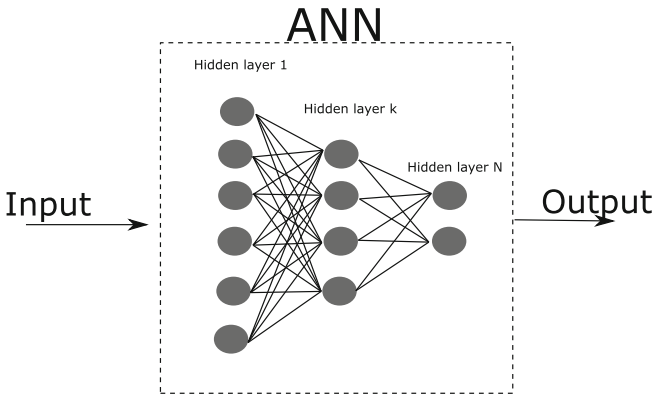


Fig. 7. ANN behavior

In this work two ANNs are used in cascade, the first one is the TDNN to model the temporal relationships of the input EEG signals. Then, the TDNN outputs are used as input to an additional PRNN. All the data elaboration steps have been performed in Matlab/Simulink R2021b. The schema of the data elaboration is shown in Fig. 8:

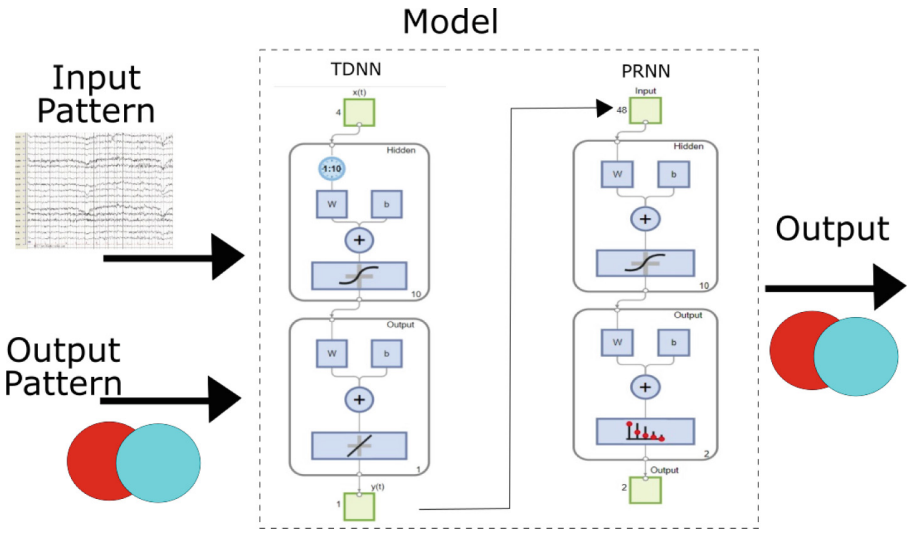


Fig. 8. Model predictive system design

3.6 Predictive Model

The aim of the model is to define an input/output function to identify a relationship between the EEG signal and the image projected during the simulation. In this paper two different type of ANN are used to perform this task. The first one used is the TDNN with 10 hidden layers and 10 samples from the 4 channels selected for the analysis $\underline{x}(t)$.

The algorithm used to calculate the ANN is the Levenberg-Marquardt backpropagation [24]. In this step a non linear function between the input pattern $\underline{x}(t)$ and output pattern $y(t)$ which is codified with value 1 or 2 must be found. The minimization cost function is the MSE between the predictive output $\hat{y}(t)$ and the observed output $y(t)$. The non linear function to be identified is:

$$\hat{y}(t) = f(x(t-1), x(t-2), ..., x(t-10)) \quad \forall t = 11, ..., T \tag{1}$$

The indexes used to evaluate the performance of the model are the MSE:

$$MSE(\hat{y}(t), y(t)) = \frac{\sum_{t=1}^{T}(\hat{y}(t) - y(t))^2}{N} \quad \forall t = 1, ..., T \tag{2}$$

and the Pearson correlation coefficient ρ:

$$\rho(\hat{y}(t), y(t)) = \frac{cov(\hat{y}(t), y(t))}{\sigma_y \sigma_{\hat{y}}} \quad \forall t = 1, ..., T \tag{3}$$

where:

– $cov(\hat{y}(t), y(t))$ is the covariance between $\hat{y}$ and y.

– $\sigma_y \sigma_{\hat{y}}$ are respectively the standard deviation of y and $\hat{y}$.

The second ANN built in cascade to the previous one is the PRNN, defined on 10 layers. The PRNN has been carried out using the scaled conjugate gradient backpropagation algorithm.

The input layer, which is the output of the TDNN has been modelled in $k = 48$ units related to an input pattern, where the i-th pattern p_i has the values resulting from:

$$p_i = \hat{y}(t(p_i)) \tag{4}$$

where

$$t(p_i) = (k+2)i + t \quad i = 0, ..., N-1 \quad t = 0, .., k-1$$

In this way the output target has been codified in two units to match the output binary target $\gamma_{i,b}$ defined on two bits $b = 1, 2$ coded as 10 for label 1 and 01 for label 2. To evaluate the PRNN performance modeled the cross-entropy loss H [25] has been chosen, computed as:

$$H = -\frac{\sum_{i=0}^{N-1} \sum_{b=1}^{2} \gamma_{i,b} \log(\hat{\gamma}_{i,b})}{N} \tag{5}$$

The objective is the minimization of the loss function which increases as the predicted probability diverges from the actual label: the lower the loss, the better the model.

4 Result

In this section,the results of the predictive model on the seven simulations of the test and validation sets are shown. The results in Table 1 show a good accuracy in both cases: test and validation for all the participants.

They are related to the performance of the predictive model composed by the two neural network:

Table 1. Accuracy of controller in test and validation set

	Test %	Validation %
P_1	74	71
P_2	86	85
P_3	72	74
P_4	81	87
P_5	72	77
P_6	67	73
P_7	85	76
All participants	62	56

The accuracy is above 67% for all cases in the test set, while in the validation set, the percentage of accuracy is more than 70% on all participants. In some participants the accuracy is low due to the noise in the data related by the hair of the participant. The electrodes are dry and may suffer by the poor contact with the scalp of the participant [22]. When trained using the half of the first simulation of all participants, the accuracy is worst, as shown in the last row of Table 1. This fact can be explained by the peculiar characteristics of the electrical signal for each participant.

Figure 10 shows the results of the TDNN in the test set in each participants. For participant P_2, the predictive and observed value are very similar, indeed according to [26] there is a strong correlation between the two output patterns $\rho = 0.75$ and a $MSE = 0.12$.

The second ANN are used to enhance the overall performance and for a better stability. In this case, using the data of the participant P_2 the accuracy of the prediction is 86% with $H = 0.3$.

The previous comments are confirmed graphically by the confusion matrix shown in Fig. 9, where the values reported are related the validation set of the participant P_2:

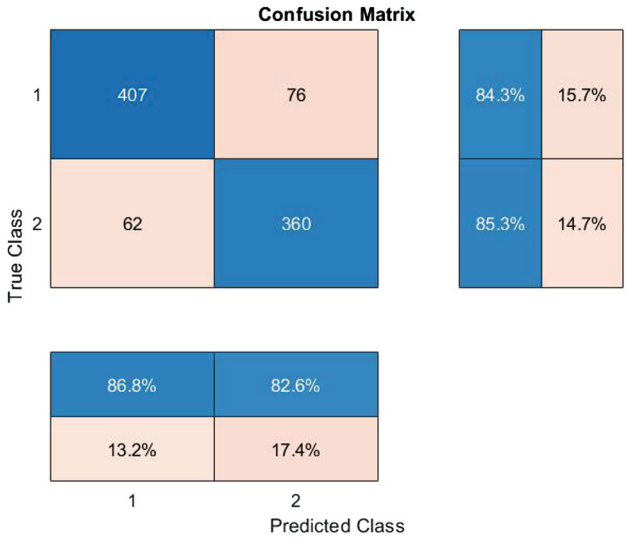


Fig. 9. Confusion Matrix in validation set for P_2

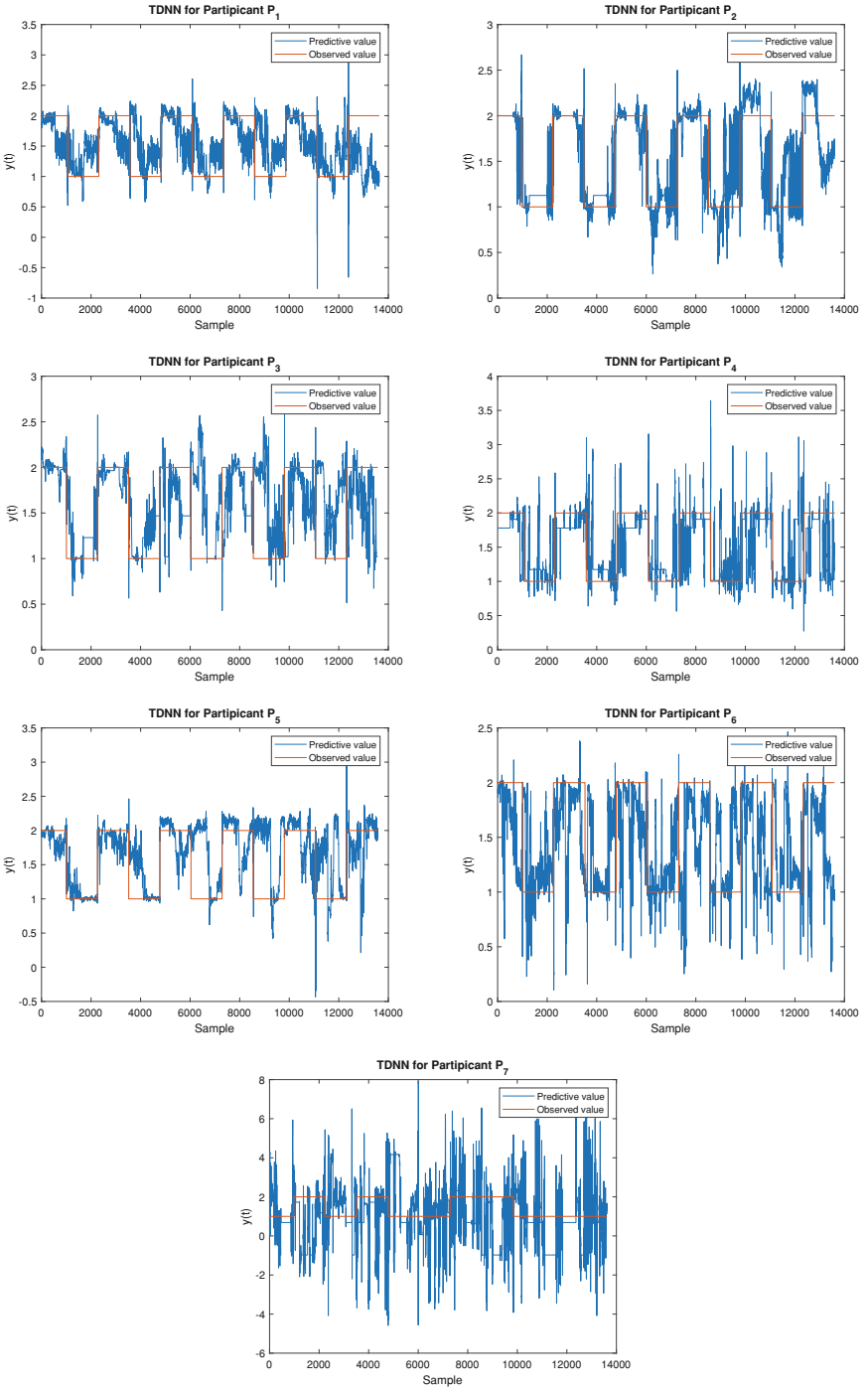


Fig. 10. Predicted versus actual trend of the images as thought by participant, from top to bottom in increasing order of participant.

5 Conclusion

The results obtained in this work are quite relevant. The model developed can predict the human thinking with an accuracy more than 70%. By the results obtained in this preliminary work, a more complex scenarios can be set up to evaluate the performance of the proposed method. The future development is related to test this experiment to a larger number of participant, and the connection with an external device to control a car in a simulated scenario or a real toy car.

Acknowledgment. This study is supported by the SysE2021 project (2021–2023), "Centre d'excellence transfrontalier pour la formation en ingénierie de systèmes" developed in the framework of the Interreg V-A France-Italie (ALCOTRA) (2014–2020), Programme de coopération transfrontalière européenne entre la France et l'Italie.

References

1. Nahavandi, S.: Industry 5.0-a human-centric solution. Sustainability **11**(16) (2019). https://www.mdpi.com/2071-1050/11/16/4371
2. Skobelev, P., Borovik, S.Y.: On the way from industry 4.0 to industry 5.0: from digital manufacturing to digital society. Industry 4.0 **2**(6), 307–311 (2017)
3. Bakay, R.: Brain-computer interfacing prospects and technical aspects. In: Schmidek and Sweet Operative Neurosurgical Techniques, pp. 1361–1372. Elsevier (2012)
4. Blum, T., Stauder, R., Euler, E., Navab, N.: Superman-like x-ray vision: towards brain-computer interfaces for medical augmented reality. In: IEEE International Symposium on Mixed and Augmented Reality (ISMAR), vol. 2012, pp. 271–272. IEEE (2012)
5. Young, B.M., Williams, J., Prabhakaran, V.: BCI-FES: could a new rehabilitation device hold fresh promise for stroke patients? Exp. Rev. Med. Devices **11**(6), 537–539 (2014)
6. Bamdad, M., Zarshenas, H., Auais, M.A.: Application of BCI systems in neurorehabilitation: a scoping review. Disab. Rehabil. Assist. Technol. **10**(5), 355–364 (2015)
7. Blankertz, B., et al.: The berlin brain-computer interface: non-medical uses of BCI technology. Front. Neurosci. 198 (2010)
8. Nijholt, A.: BCI for games: a 'State of the Art' survey. In: Stevens, S.M., Saldamarco, S.J. (eds.) ICEC 2008. LNCS, vol. 5309, pp. 225–228. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89222-9_29
9. Bi, L., Fan, X.-A., Liu, Y.: EEG-based brain-controlled mobile robots: a survey. IEEE Trans. Human Mach. Syst. **43**(2), 161–176 (2013)
10. Jahan, M., Khan, Y.U., Sharma, B.B.: Classification of EEG signals based on imaginary movement of right and left hand wrist. In: 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), pp. 193–196 (2014)
11. Hashimoto, Y., Ushiba, J.: EEG-based classification of imaginary left and right foot movements using beta rebound. Clin. Neurophysiol. **124**(11), 2153–2160 (2013)

12. Sakhavi, S., Guan, C., Yan, S.: Learning temporal information for brain-computer interface using convolutional neural networks. IEEE Trans. Neural Netw. Learn. Syst. **29**(11), 5619–5629 (2018)
13. Zero, E., Bersani, C., Sacile, R.: EEG based BCI system for driver's arm movements identification. In: Proceedings of the Automation, Robotics & Communications for Industry 4.0, vol. 77 (2021)
14. Umamaheswari, J., Akila, A.: An enhanced human speech emotion recognition using hybrid of PRNN and KNN. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 177–183 (2019)
15. Zero, E., Bersani, C., Zero, L., Sacile, R.: Towards real-time monitoring of fear in driving sessions. IFAC-PapersOnLine **52**(19), 299–304 (2019)
16. Binias, B., Myszor, D., Cyran, K.A.: A machine learning approach to the detection of pilot's reaction to unexpected events based on EEG signals. In: Computational Intelligence and Neuroscience, vol. 2018 (2018)
17. Yin, E., Zhou, Z., Jiang, J., Yu, Y., Hu, D.: A dynamically optimized SSVEP brain-computer interface (BCI) speller. IEEE Trans. Biomed. Eng. **62**(6), 1447–1456 (2014)
18. Cao, L., Li, J., Ji, H., Jiang, C.: A hybrid brain computer interface system based on the neurophysiological protocol and brain-actuated switch for wheelchair control. J. Neurosci. Methods **229**, 33–43 (2014)
19. Zheng, B.C., Li, Q., Xiang, Y.X., Hernandez, J., Mann, S.: Brain and visual controlled mobility scooter
20. Razoumnikova, O.M.: Functional organization of different brain areas during convergent and divergent thinking: an EEG investigation. Cognit. Brain Res. **10**(1–2), 11–18 (2000)
21. Singh, V., Veer, K., Sharma, R., Kumar, S.: Comparative study of fir and IIR filters for the removal of 50 Hz noise from EEG signal. Int. J. Biomed. Eng. Technol. **22**(3), 250–257 (2016)
22. Zero, E., Bersani, C., Sacile, R.: Identification of brain electrical activity related to head yaw rotations. Sensors **21**(10), 3345 (2021)
23. Dongare, A., Kharde, R., Kachare, A.D., et al.: Introduction to artificial neural network. Int. J. Eng. Innovat. Technol. **2**(1), 189–194 (2012)
24. Lv, C., et al.: Levenberg-marquardt backpropagation training of multilayer neural networks for state estimation of a safety-critical cyber-physical system. IEEE Trans. Indust. Inform. **14**(8), 3436–3446 (2017)
25. Feng, D., Chen, L., Chen, P.: Intention recognition of upper limb movement on electroencephalogram signal based on CSP-CNN. In: 2021 5th International Conference on Robotics and Automation Sciences (ICRAS), pp. 267–271. IEEE (2021)
26. Cohen, L.H.: Life Events and Psychological Functioning: Theoretical and Methodological Issues, vol. 90. SAGE Publications, Incorporated (1988)

Tactile-Based Human-Robot Collaboration: A Performance Analysis

Francesco Grella[1(✉)], Roberto Canale[1], Francesco Giovinazzo[1], Alessandro Albini[2], and Giorgio Cannata[1]

[1] Universitá degli Studi di Genova, Genova, Italy
Francesco.Grella@edu.unige.it
[2] Oxford Robotics Institute, Oxford, UK

Abstract. In this paper we present a performance analysis of a perceptual architecture for industrial robots based on tactile sensing. We have developed a human-robot interface based on sensing devices which can be fixed on the robot body or the gripper. The devices are cylinder-shaped handles covered with tactile sensors, which can be intuitively grasped by an operator. Tactile data are fed into a neural network in order to recognize human touch during grasp, thus providing an enabling command for the control system. We provide an inference time comparison between two computing architectures: a desktop workstation with an high-performance GPU and an embedded solution based on the NVIDIA Jetson Nano board. We also compare inference time obtained from three instances of the neural network, compiled with three different engines: Keras, TensorRT floating-point 16 and TensorRT floating-point 32, showing that numerically optimized models allow to perform inference also on the embedded board without violating timing constraints imposed by data acquisition. Moreover, we assess the robustness of touch recognition when the user is wearing work gloves, showing that the difference from the bare-hand case is negligible.

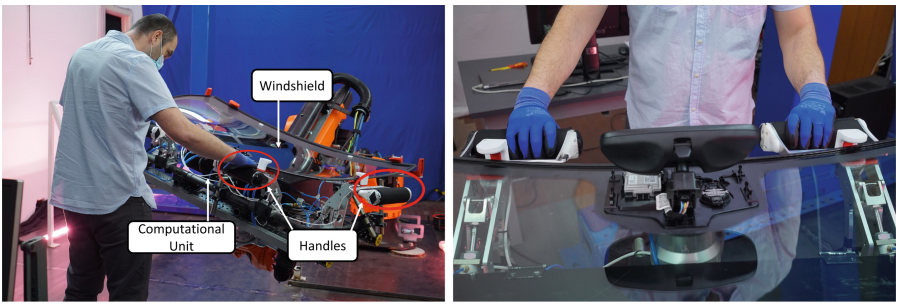
Keywords: Tactile sensing · Human-robot interaction · Deep learning

1 Introduction

Industry 4.0 paradigm [1] foresees the co-existence of humans and robots in applied manufacturing scenarios. Human's experience and dexterity should be leveraged to handle new situations and solve complex problems, while robot's strength, high speed and accuracy should be used to accomplish monotonous, dangerous and repetitive tasks. Compared to more traditional manufacturing layouts which expect the robot to be confined within a restricted workspace, *human-robot collaboration* (HRC) aims towards the realization of tasks performed in a shared workspace. This novel working configuration allows to define

Supported by the European Community's Framework Programme Horizon 2020 (grant agreement number 820767, project CoLLaboratE).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 437–445, 2023.
https://doi.org/10.1007/978-3-031-16281-7_41



(a) (b)

Fig. 1. Two views of the use-case scenario. Image (a) shows the gripper which holds the windshield. Data acquisition is performed by devices contained in the black box below the vacuum which holds the glass. In image (b) the two front handles and the work gloves of the operator are clearly visible. The model is similar to the glove used for robustness experiment.

and perform operations that can exploit capabilities both from the human and the robot, but also raises challenges in terms of safety and communication between the operator and the machine. The analysis described in this paper is contextualized in the car windshield assembly use-case of the CoLLaboratE H2020 [2] project. We designed a cell to achieve HRC by means of physical Human-Robot Interaction (pHRI), a specific class of interaction situation in which an operator is expected to make contact on one or more points of the robot body. In the depicted scenario (presented in Fig. 1), the industrial manipulator is used as a weightlifting system which absorbs the whole effort of moving an high-inertia object around the workspace. The capability of effortlessly moving a windshield by hand also improves the flexibility of maintenance operations, since the user can now tailor and adjust the position and orientation of the object according to his/her needs. If the inspection and manufacturing operations would occur on a fixed object, the worker should instead adapt to the position of the object with the risk of being penalized by proper anatomic characteristics such as height.

In order to be guided by an operator, the robot motion must be governed by an admittance controller which takes as input an external force/torque and outputs a proper velocity command which makes the robot follow the applied wrench. Such motion behavior anyway does not distinguish who or what is actually applying the force: if the operator hits the robot accidentally an undesired and dangerous movement will be performed. Also an object inattentively laid on the payload or the gripper could cause a constant force applied on the robot, thus a constant velocity which can bring the robot to collide with the ground, other equipment and people. These considerations justify and contextualize the need of a way to discriminate between accidental and voluntary contact. In this scenario human contact is used as a novel communication interface through which the human can directly manifest his/her intention to interact with the robot by

classifying the data obtained from the tactile sensors. Voluntary interaction is detected when the shape of a human hand is recognized by the classifier, meaning that the operator is deliberately grasping a handle in order to move the robot.

Endowing an industrial manipulator with an external sensing architecture whose feedback actively affects motion control yields implementation issues which need specific design choices. In pHRI one major requirement is safety, which in reactive control architectures also depends on system's response. In the proposed application the activation/deactivation of the robot motion based on hand detection should not introduce latency to the control system, so data processing should be performed within the time window of each new acquisition step. Thus, given the rate bottleneck of the sensors and knowing that inference from deep learning models is a computationally expensive processing task, proper hardware should be used for online inference. In literature, pHRI is addressed with different sensing architectures. Authors in [3,4] address physical interaction by means of admittance control strategies relying only on the feedback provided by joint level torque sensors on small-sized robots. These solutions are unfeasible when dealing with heavy industrial manipulators, since joint level sensors are not a common design choice. Moreover heavy-duty robots lack the necessary sensing capability to detect and localize a light contact due to the weight of their links.

To endow industrial manipulators with physically interactive behaviors purely reactive compliance-based approaches could not be enough to ensure safety standards described in works such as [5]. In [6] a survey on human-robot collaboration in industrial settings is presented, however only few works fall in the area of 'shared-payload' tasks like the one described throughout this work. In [7] an industrial robot as flexible weightlifting system is presented, but no direct measurement feedback between the machine and the operator is considered. Therefore a distributed tactile interface providing a higher level abstraction about the tactile information should be leveraged. In [8] tactile image processing is used to semantically interpret grasp poses to produce discrete motions on a lightweight manipulator. Among the cited papers authors anyway do not consider use-case scenario and therefore no analysis on implementation details is provided. Moreover, in [8] which addresses human hand detection no considerations are provided with respect to robustness in presence of protective clothing such as gloves, which are mandatory in manufacturing environments. An application based on tactile sensing to detect user grasp should consider the fact that the operator is very likely to wear protective gloves which can alter the sensors response. To the best of our knowledge, no such analysis is available in literature. Thus, the main contributions of this paper are the following:

- ◇ a performance analysis of the learning model in terms of inference time. In particular we carried out a comparative analysis between three instances of the same model: a Keras [9] model and two models compiled with the optimization framework *TensorRT* [10] with different floating point precision. The tests have been iterated on two different computational architectures based on a RTX 3070 GPU and a Jetson Nano board.
- ◇ a robustness analysis on hand detection while the user is wearing a protective glove, to assess its impact on the performance of the classifier.

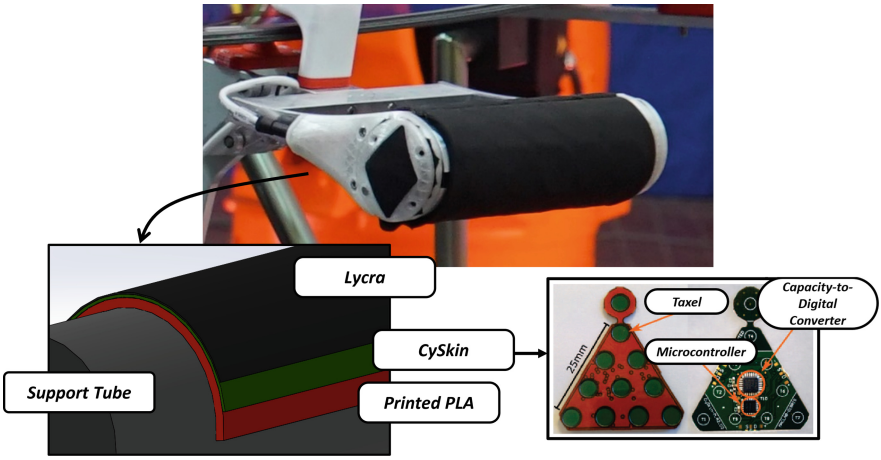


Fig. 2. Detail on the device structure and CySkin architecture.

The rest of the paper is organized as follows:

Section 2 provides an overview of the sensing architecture implemented for the use-case and details regarding the fine-tuning procedure we adopted to improve accuracy performance. Section 3 describes inference time evaluation, while Sect. 4 presents results related to robustness analysis. In Sect. 5 we provide our conclusions as well as insights for an extension of the research.

2 Voluntary Interaction Detection Architecture

This paper presents a performance analysis of an architecture which allows a human operator to physically interact with an industrial robot. The system architecture and methodology are already presented in detail in [11,12], so it will only be briefly reviewed in this section. Our pipeline builds a 2D representation identified as *tactile image* from raw tactile data, which is then processed by a binary classifier based on a convolutional neural network which outputs the probability of both *hand* and *non-hand* classes. The output value of the *hand* class is then used to enable robot motion control. The device designed for this application is a cylindrical handle covered by distributed tactile sensors arranged in flexible patches which allow to cover curved surfaces. The handle is covered with the CySkin technology, presented in [13]. It consist of an artificial skin composed of distributed tactile sensing elements (i.e. *taxels*) measuring the pressure distribution generated by the physical interaction. The skin is composed by flexible interconnected modules that can be cut and adapted to curved surfaces. Each modules can contain up to 11 capacitive pressure transducers. New data are sampled at a rate 20 Hz which represents the main timing constraint for our application. It's desirable to meet an inference time which is below 0.05s in order to classify voluntary interaction within the acquisition step. Figure 2 shows the

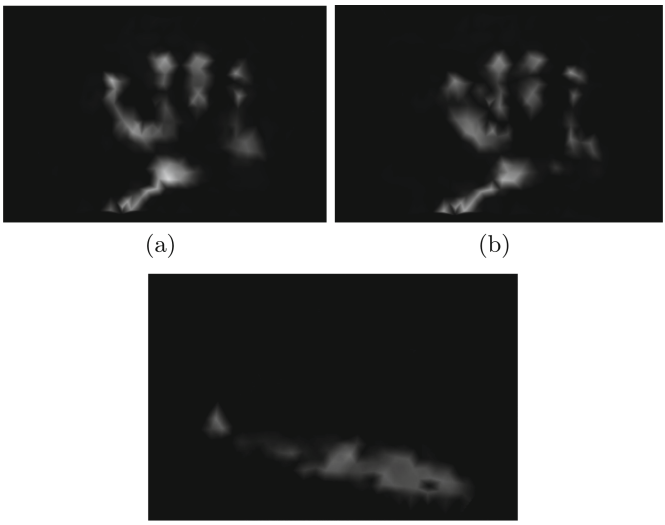


Fig. 3. Three instances of tactile images during interaction events: (a) grasp with bare hand, (b) grasp with working gloves, (c) forearm contact. It can be noticed that when the user is wearing the glove the region between the index and middle fingers is detected as pressed due to the thickness added by the tissue.

structure of the handle and the CySkin technology. Given the discrete set of pressure sensors, a single sensing element is referred as *taxel* with an associated pressure response. By following a geometric pipeline described in [14] the set of pressure values is interpolated over a triangular mesh approximating the covered surface. A further flattening step leads to a *2D* surface which is transformed in a greyscale image. Figure 3 shows three examples of tactile images acquired during the experiments for the presented analysis. At each acquisition step a tactile image allows to acquire information on the contact shape. We use a convolutional neural network (*HandsNet*) presented in [11] in order to predict the probability of having a human hand touching the device given the image. The training set was acquired from human subjects touching and grasping the skin-covered forearm of a robot. In order to improve accuracy on images acquired from the handle we performed a fine-tuning of the network which is presented in detail in [12]. The training procedure led to and accuracy of 98.08% and 94.61% on the validation and test sets respectively. The work depicted in [12] also presents an early performance analysis of the detection system on the Jetson Nano board. Results showed a slightly better performance of *fp16* against *fp32*, but the analysis was more conceived as a proof-of-concept for the system and has several differences from the study proposed in this article. In particular a lower amount of data was used and no online acquisition was performed, but frames were taken from

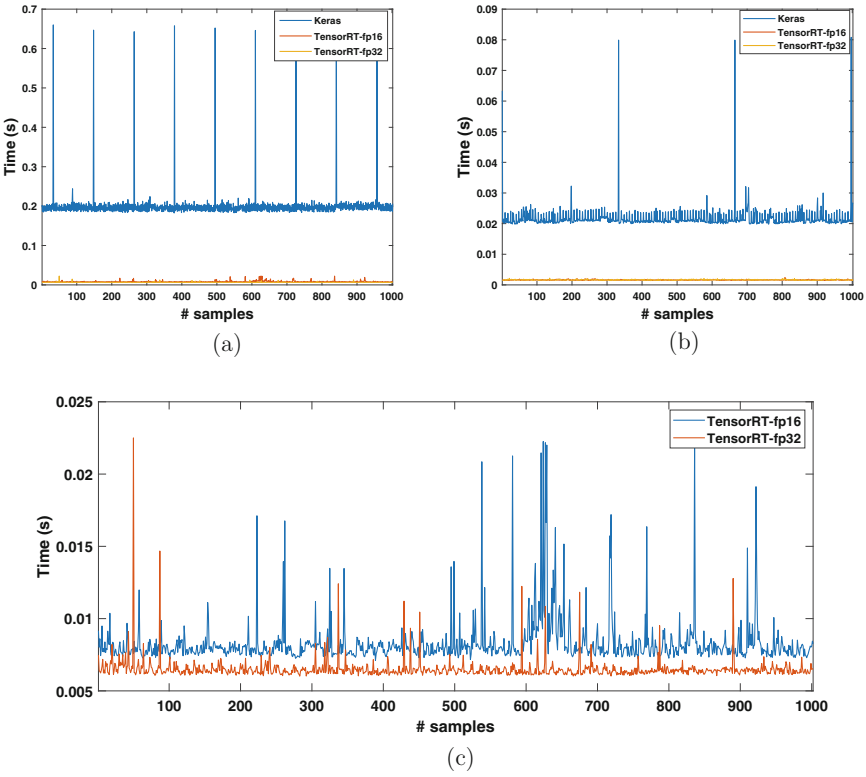


Fig. 4. Inference time plots on 1000 samples. (a) shows the inference time obtained on Jetson Nano, (b) the inference time obtained on RTX 3070. (c) is an insight on fp16 and fp32 obtained on Jetson Nano experiments

a pre-recorded dataset. Moreover the timing analysis also included preliminary processing steps such as image resizing for the input layer of the network, since it is focused on the architecture feasibility more than pure inference performances.

3 Inference Time Performance Evaluation

We propose a comparison between two devices: a desktop workstation with a *RTX 3070 GPU* and a *Jetson Nano* board, both manufactured by Nvidia. Assuming that data are acquired by devices which are placed on the robot gripper, the adoption of a workstation for computing inference requires a more complex communication architecture since the size and weight of the machine do not allow direct placement on the robot. A local network should be used to send data from the on-board sensors to the computing device and then the prediction back to the robot controller, eventually introducing a source of latency. On the other hand, having a small GPU board placed directly on the gripper allows to perform data acquisition and model inference on the same device, leaving

Table 1. Results of the inference time analysis, expressed in seconds (s).

	Keras			TensorRT (fp16)			TensorRT (fp32)		
	Mean	Max	Std	Mean	Max	Std	Mean	Max	Std
RTX 3070	0.0216	0.0808	0.0038	0.0016	0.0024	9.8×10^{-5}	0.0016	0.0022	1.13×10^{-4}
Jetson Nano	0.1991	0.6786	0.0438	0.0083	0.0224	0.0017	0.0065	0.0225	8.01×10^{-4}

only one communication channel towards the controller. We have performed an inference time analysis by running the model on real data acquired by CySkin in order to assess the feasibility of both solutions. Since the presented analysis is independent from the robot motion, we have used a single sensorized handle fixed on a table. Since this study presents an analysis purely focused on detection performances, data have been acquired from a single subject interacting with the device, and a set of 1000 frames has been considered for each experiment. An in-depth user study which also involves robot motion performances is currently being performed. We also have considered three different instances of HandsNet for each architecture as a complement for our analysis. We have collected inference time from a plain Keras model directly obtained by the framework used for building the network and from two models obtained by compilation with TensorRT (TRT), a toolchain which provides optimization for GPU architectures. The difference between the two optimized models is the floating-point precision used for representing the network parameters: here we discuss the behavior of *floating-point 16* and *floating-point 32*.

Figure 4 (a) and (b) and Table 1 show that the Keras model exhibits poor time performance compared to TRT models on both devices. Only on the workstation the time constraint is met by observing the mean inference time. Anyway, it must be noticed that peaks reaching $0.0808s$ are detected, so there is no guarantee that inference can be always performed within the same acquisition step. On both architectures TRT models have obtained acceptable results in terms of inference time. On RTX 3070 the difference in performance between fp16 and fp32 is negligible, instead on the Jetson board the fp32 model has a mean inference time improvement of 1.8 ms. The fp32 model on Jetson also shows lower standard deviation with respect to the fp16 counterpart. The better quantitative performance and stability of the fp32 model is also shown in Fig. 4 (c).

4 Inference Accuracy Robustness

The proposed use-case scenario is set in an industrial environment and is focused on physical interaction between the robot and a specialized worker. Voluntary interaction detection is performed by classifying a human hand shape from a tactile image when the human is grasping the devices described in Sect. 2. In assembly departments there are strict safety policies which the workers must follow, especially regarding protective clothing. In the described scenario the operator is expected to grasp the handles while wearing protective gloves, like

shown in Fig. 1. Since the neural network was originally trained with bare hand grasps, we provide an assessment of its classification performance when grasp is performed while wearing gloves. We acquired data from a user grasping one fixed handle and holding the grasp for two minutes while randomly rotating the wrist to emulate the small motions which occur while the robot moves. We analyzed the probability value of the hand class in terms of mean, minimum value and standard deviation. As shown in Table 2 wearing a work glove has a small impact on the prediction value. In particular the presence of the glove slightly increases fluctuations as can be seen by observing the standard deviation and the minimum value, but in terms of classifications these values are negligible even by choosing a tight threshold of 0.9 to enable the controller.

Table 2. Results of comparison between hand class probability in bare hand grasp and work glove grasp. Reported results represent probability values, thus they are dimensionless and limited between 0 and 1.

	Mean	Min	Std
Bare hand	1.00	0.99	5.14×10^{-6}
Work glove	0.99	0.95	0.0019

5 Conclusions

In this paper we have presented a performance analysis for a tactile-based sensing architecture designed for physical human-robot interaction in an industrial scenario. We have evaluated the impact of the computational architecture and deep learning model optimization in terms of inference time, showing that an embedded board with a low-cost GPU is capable to provide a prediction for each acquired image without exceeding the sampling period of $0.05s$. Both fp16 and fp32 models allow to obtain performances on the Jetson board which are very close to those observed on the RTX 3070. This shows that model optimization could enable deep learning model inference in applications with tight space or power consumption constraints. We have also tested the robustness of our hand classification model with respect to grasp performed while wearing working gloves. We have obtained results showing that this common safety practice in industrial environment does not affect voluntary interaction detection. Further steps in our research will involve performance evaluation during operations on the full use-case scenario, as well as a user study on multiple subjects. Both studies are currently under way.

References

1. Xu, L.D., Xu, E.L., Li, L.: Industry 4.0: state of the art and future trends. Int. J. Prod. Res. **56**(8), 2941–2962 (2018)
2. Collaborate Project. https://collaborate-project.eu/

3. Magrini, E., Flacco, F., De Luca, A.: Control of generalized contact motion and force in physical human-robot interaction. In: 2015 IEEE International Conference on Robotics and Automation (ICRA), pp. 2298–2304 (2015)
4. Keemink, A.Q., van der Kooij, H., Stienen, A.H.: Admittance control for physical human-robot interaction. Int. J. Robot. Res. **37**(11), 1421–1444 (2018)
5. Bragança, S., Costa, E., Castellucci, I., Arezes, P.M.: A brief overview of the use of collaborative robots in Industry 4.0: human role and safety. In: Arezes, P.M., et al. (eds.) Occupational and Environmental Safety and Health. SSDC, vol. 202, pp. 641–650. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-14730-3_68
6. Villani, V., Pini, F., Leali, F., Secchi, C.: Survey on human-robot collaboration in industrial settings: safety, intuitive interfaces and applications. Mechatronics **55**, 248–266 (2018)
7. Gopinath, V., Johansen, K.: Risk assessment process for collaborative assembly – a job safety analysis approach. In: Procedia CIRP, vol. 44, pp. 199–203 (2016). 6th CIRP Conference on Assembly Technologies and Systems (CATS).
8. Li, T., Zheng, K., Liu, J., Jia, X., Feng, J.: Research on operation intention based on flexible tactile sensing handle. IEEE Access **9**, 12362–12373 (2021)
9. Chollet, F., et al.: Keras (2015)
10. Nvidia, tensorrt. https://developer.nvidia.com/tensorrt
11. Albini, A., Cannata, G.: Pressure distribution classification and segmentation of human hands in contact with the robot body. Int. J. Robot. Res. **39**(6), 668–687 (2020)
12. Grella, F., et al.: A tactile sensor-based architecture for collaborative assembly tasks with heavy-duty robots. In: 2021 20th International Conference on Advanced Robotics (ICAR), pp. 1030–1035 (2021)
13. Maggiali, M., Cannata, G., Maiolino, P., Metta, G., Randazzo, M., Sandini, G.: Embedded distributed capacitive tactile sensor. Mechatronics **7** (2008)
14. Cannata, G., Denei, S., Mastrogiovanni, F.: On internal models for representing tactile information. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1108–1113 (2010)

Applications and Challenges in the Use of 5G in Production and Logistics

Wireless Communication Technologies in Smart Factories

Dieter Lutzmayr(✉) and Manfred Pauritsch

CAMPUS 02 University of Applied Sciences, 8010 Graz, Austria
dieter.lutzmayr@campus02.at

Abstract. 5G offers high speed and low latency while providing better network coverage and better handover than wireless LAN. This makes it a good choice for any kind of remote control and remote computing, where the on-site devices just have to process sensors and actuators and the main tasks can be conducted remotely in real time. Therefore, the sensor devices can be rather small, light and with low computation power, which may enable energy harvesting or allows to be carried by humans. Especially when considering data streams like video, the effort to process this huge amount of data in real-time is high. If this can be done either by humans that do not have to be at the location of the action or by servers that have access to the whole factory data set, a significant improvement can be achieved in terms of work augmentation. One use case is uninterrupted real-time remote controlling, where the person in charge does not have to move physically, like with UAVs for transportation, inventory or monitoring purposes. Inversely people in the factory can be augmented for working, training or maintenance purposes. Person and head movements are processed on a server farm to create an augmentation image that adds an information overlay to the real world on data glasses, giving instant access to individually selected resources with big data from the whole plant.

Keywords: Augmented · Real-time · 5G · Remote control · Remote rendering

1 Introduction

The idea of digitally mapping production to manage the entire life cycle more efficiently, from planning to maintenance, has been around for decades. In the 1980s, this approach emerged under the term Computer Integrated Manufacturing, but was not feasible with the technology of the time. Today, the term "digital twin" is often used. The physical object of production is mapped in a digital model and information is exchanged bidirectionally via a data interface. Due to today's technological possibilities in software, hardware and data transmission, many industrial companies are also working on the vision of the digital twin. Magna Steyr Fahrzeugtechnik also has a strategy for a "smart factory" [1].

Magna Steyr Fahrzeugtechnik is a production plant in the industrial sector of automotive with an area of half a square kilometer and a maximum extent of nearly one

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 449–457, 2023.
https://doi.org/10.1007/978-3-031-16281-7_42

kilometer spanning over one main street. Magna has the vision to add flexibility to production sites by implementing only wireless communication and removing all wired network connections. The objective is to overcome following issues:

1. The development to a smart factory will increase the number of mobile and connected devices such that WLAN access points are not able to manage the number of devices
2. Even today the installation and maintenance of WLAN access points is a significant cost factor both in cabling hardware and working hours.

Furthermore, wireless communications will allow rapid reconfiguration of future factories, necessary to adapt to new products and processes.

In this paper, three use cases are presented for which Magna and their academic partner Campus 02 UAS have evaluated the usage of wireless communication technologies. The paper is structured as follows: After an introduction to the general requirements and the description of the test infrastructure in Sect. 2, use cases are presented in Sect. 3. Both results and hindering factors are described in Sect. 4, finally a conclusion is given in Sect. 5.

2 Industrial Requirements and Test Infrastructure

To overcome distance limitations and security issues and allow a significant number of systems at a time and therefore make remote controlling useful for industry applications, there is the need for a wireless system that offers certain features:

- Enough power for distances up to 1 km in free air
- Full coverage inside, outside also under conditions of distorted environments
- Automatic handover between base stations
- Connection to long-distance communication means, especially the internet
- Availability of cryptographic methods to protect data communication
- Defined Quality of service to provide a certain level of performance
- Single sender with multiple receivers that both can change dynamically
- Possibility of private networks, allowing full control over the system on site

Remote controlling can be enhanced even further if the controller is no human person but some IT devices. Typically, this will be a server farm running complex machine learning or artificial intelligence algorithms, namely a cloud solution on the internet or on the edge of the 5G network.

2.1 Test Infrastructure

The typical wireless infrastructure with currently available technologies relies on Wireless LAN (WLAN) with additional Global System for Mobile Communications (GSM) cellphones that are data enhanced with Long Term Evolution (LTE). Both have their own disadvantages concerning the above-mentioned three issues: WLAN is limited to a certain number of devices per access point, for large plants one need a lot of access

points while there still will be some areas without radio reception, and seamless handovers are hard to manage especially for faster moving devices. While some of those issues can be circumvented by just moving to another position, which is easy for humans or human operated vehicles, this is not so easy for automated systems. Certain improvements like 5 GHz WLAN have been eaten up very fast by the rapid increasing number of devices, so also all guest networks had to be disabled in the plant. LTE seems to be a solution for most of the requirements but has the big advantage that all data has to be routed through the infrastructure of the telecommunication provider, adding latency and further bottlenecks, reliability and availability problems as well as security issues.

There are further requests to the wireless infrastructure that cannot be fulfilled very well by the current technologies, namely robust and accurate localisation and especially real-time operation. Therefore, current navigation systems for automated guided vehicles as well as for manual forklifts with automatic position detection and goods recognition rely on onboard computing power. After an initial setup each vehicle carries its own map of the whole site, at least the software exchanges data about changes in the environment between the individual vehicles.

While WLAN is not ideally suitable for big distances and a lot of moving clients, public infrastructures like GSM and its faster data communication enhancements like Universal Mobile Telecommunications System (UMTS) 3G and LTE (4G) do not fulfill the quality and privacy aspects for industrial use cases. This changed with 5G, where some main aspects are the increase of connected devices per squarementer, quality of service routing and the possibility to set up campus networks, while still maintaining the advantages of mobile networks.

The following technologies are selected to evaluate the performance, the possible coverage and the overall usability of current wireless technologies:

- Wireless LAN (WLAN) with standards 802.11ac (Wi-Fi 5) and 802.ax (Wi-Fi 6), since currently there are no devices available with 802.11be (Wi-Fi 7) for mobile systems using current technology as a comparison basis
- 5G campus installations provided by two different providers at Magna and at CAMPUS 02 UAS
- Ultra Wideband (UWB) for precision localization
- Local Range Wide Area Network (LoRaWAN) for ultra low power embedded devices.

The Magna site is completely WLAN-covered with current Cisco hardware, offering both 2.45 GHz and 5 GHz access points. Magna also installed a test bed for 5G at the production site, set up by the provider Telekom Austria. On the client side most devices offer Wi-Fi reception that is used as the basis for Wireless LAN tests, while 5G uses external LAN-connected 5G-Routers, namely the FRITZbox 6850 and the Nokia FastMile 5G Gateway. The requested embedded 5G modules Quectel RM50xQ and Telit FN980 are still not available.

The other infrastructure is located on the Campus 02 site, where the on-site installation with components from Nokia is directly connected via glass fiber to the datacenter of Citycom Graz, a city-owned telecom provider focused on high quality business services. Hence, testing remote cloud services like online remote rendering and voice recognition is possible.

While at Magna all necessary remote computing is conducted on the edge on their own IT cloud infrastructure with software implemented by a Magna subdivision, the CAMPUS 02 UAS test site is connected to Microsoft Azure to make use of the offered features.

Additional tests included UWB and LoRaWAN. The LoRaWAN setup consists of a Laird Sentrius Gateway RG 186 equipped with Wi-Fi, Bluetooth and Ethernet as the base station and ST B-L072Z-LRWAN1 LoRaWAN Discovery Kits as mobile devices. For UWB the Mobile Knowledge Evaluation Kit RTLS [2] is used, which is based on the NXP Trimension SR150 and SR040 chips.

3 Use Cases and Evaluation Criteria

The first use case is logistics within the production site, the second use case is the usage of augmented system control and the third use case is augmented rendering. They are presented in the following sections.

3.1 Use Case 1: Transport and Order Picking of Goods

A very important factor is flexibility in the production line. Because of batch size 1 in conjunction with individualization, products undergo different production steps and the products rapidly change. For providing these items at production line the number of transports increases as not anymore big carriers e. g. pallets with identical items but small carriers e. g. boxes are transported. Also, batch size 1 relies on so-called sets of items to be provided just in sequence (JIS) providing necessary goods for each individual product to be assembled or produced. To minimize the error rate for transports and order picking processes, regular identification of goods by usually using bar codes is conducted. Both processes are described briefly in the following paragraphs.

For the transport process at the source of the transport both location and good to be transported are identified via scanning barcodes. This is done by the driver/AGV being equipped with a mobile/installed bar code scanner, which in case of manual transport is connected to the terminal. After having conducted the transport again the good is scanned and is coupled with the source also being scanned.

For order picking processes in batch size 1 production lines supermarkets are located at several local places within the production line to pick necessary items, combine them to a set and put them in a box. Within this process every picked item has to be identified and has to be coupled with the corresponding box. This process is due to the high variety of items for individualized products highly error prone. Furthermore, especially in production lines with a fixed cycle time missing parts result in complex processes of removing and inserting goods into the production line.

To ensure availability of connectivity at all locations in the smart factory coverage with wireless communication is to be ensured, which is consequently one of the measurement criteria for this use case.

3.2 Use Case 2: Augmented Control

The remote control of home appliances and hobby systems needs only a quite small bandwidth to send commands from the controller to the controlled device. Examples are remote control for audio or video systems or toys that use modulated infrared light or remote control for model cars, planes or helicopters that use radio control in the open bands of 27 or 40 MHz. To control, the operator has to have direct sight to the controlled object, since there is no feedback channel. Beside of the issue of having only limited free bands there is no means to protect the system from distortions or eavesdropping, rendering it useless for industrial applications. With modern drones, a new type of remote control transmission in the 2.4 GHz band has appeared that offers a video and data backward channel. At most commercially available systems a smartphone can be connected to the controller and will be used to display and record live video streams or high-resolution pictures. While this significantly improves the quality of remote controlling, the distance between sender and receiver is limited by the power of the sender that is limited by governmental restriction. In addition, the number of systems at a certain location is limited by the available bandwidth.

Augmented control has two different use cases at Magna. Through increasing amount wireless communication partners (see above) leads to a new issue: All of those mobile systems have to be placed in position and/or taught concerning functionality, setup and maintenance. For this task, still people have to move to the corresponding place, as well as for the fixing of problems. For such tasks teleoperation will be useful, but only if there is visual real-time feedback as well as real-time control. For comfortable usage of this technology short latency and a high throughput is necessary.

Especially the last two years with the pandemic and restricted possibilities to travel to other countries Magna has the necessarity to train new employees at sites in China, which is in ramp-up phase. Augmented training for workers in a Chinese plant is of a high value to enhance success of training. Experienced workers give direct visual feedback to the remote workers that wear Microsoft HoloLens 2 AR glasses. Through visual augmentation, also the language barrier can be overcome.

3.3 Use Case 3: Augmented Rendering

Modern industry plants include many systems that contain a big number of sensors and actuators as well as they are complex in terms of setup and maintenance. Currently the personnel carries printed information or electronic devices, and with every step, they have to read first, do the work and then check whether it was successful or not. This effort can be reduced significantly, if all the information is available directly and digitally. One approach to overcome this efficiently is Remote Rendering. By using Remote Rendering high quality 3D content and interactive information are displayed on Mixed-Reality-Devices e.g. Microsoft HoloLenses. This is already well known in massive multiplayer online games (MMOGs), where many clients interactively collaborate to give a highly realistic real-time experience. Research shows that for such scenarios remote rendering can be a good option and will allow game participation with inexpensive low power hardware [4]. This also holds true for mobile devices in industrial environments, since

most of them are battery powered, this battery should last at least one shift, which is likely to last 8 h of nearly continuous work.

The key component to a successful integration is the availability of data as well as the ability to display it on a mobile device. Starting with Google glasses the offering of augmented reality glasses began, where the workers wear special glasses that mirror the requested information onto the inner side of the glass or use semi-transparent glasses. To be able to connect the information to the environment, e.g. mark objects/places, show how to apply a tool or show status information connected to items, movements of persons and especially their heads have to be tracked in real-time. In addition, rendering must be applied in real-time, connecting the plant data to the visual image. The amount of data necessary to accomplish remote rendering for industrial scenarios is rather big, as well as there is a need for computing power to create the images in real-time. For larger industry plants, it is impossible to store all data on a local device, not to mention updates and real-time sensor data. Remote rendering mechanisms have been presented and evaluated already some years ago [3], all of them share the same necessities: Fast connection with high data rate and a low latency.

When using remote rendering on powerful edge or cloud servers, also other algorithms and mechanisms can be used for further information gaining and control. With this possibility new opportunities arise as voice recognition, where the augmented reality assisted workers can control the information by natural voice instead of gestures or tapping. Also live streams from mobile camera(s) and microphone(s) can be used for analysis with machine learning or AI algorithms, which in return will offer further process optimization possibilities.

At Magna already the whole plant is available as a digital model in terms of mechanical design. Since also most of the documentation is available electronically, both parts were linked together. Therefore, it is possible to use remote rendering to add personalized information to the environment for which Microsoft HoloLens 2 AR glasses are used.

3.4 Evaluation Criteria

For all of the use cases criteria are derived:

- Use Case 1:
  - Received signal strength of LoRa devices to test the ability to send status information from different positions
  - Position estimation based on triangulation to base stations
  - Fine location detection using UWB
- Use Case 2: Measurement of signal strength, latency and throughput of data transfer to a central server to evaluate whether these parameters are sufficient for a full remote control including server based navigation instead of onboard navigation, both with WLAN and 5G.

- Use Case 3: Frame rate and latency as well as the perceived quality of transmitted video streams on site and with Internet connection to the Chinese Magna plant in Zhenjiang.

4 Results and Hindering Factors

UWB is very promising for exact positioning with an accuracy in the range of millimeters, but it also needs a local installation with at least two so called anchors. LoRa shows a very good performance in terms of coverage and distance as well as battery usage and lifetime, but due to the nature of the system, only very small chunks of data may be transmitted. The data rate is also limited by the fact that the airtime is strictly limited. Sigfox as a public LoRa network already uses some of the bandwidth. For industry the usage of such networks is not desired because of security issues, no information may leave the production plant. In the end, most of the experiments have been a comparison of WLAN and 5G.

Overall, 5G showed the best wireless network performance results, with the exception that only a few 5G devices could be used. Therefore, the end devices that were used are rather clumsy and need AC power. Another issue is the availability of Virtual Reality (VR) and Augmented Reality (AR) headsets. While for static setups, where the person stays in a certain location, the non-transparent VR glasses can be used, this is not possible for moving people. Not only because of the latency that is dangerous in many cases, but also because of the technology, that uses infrared senders for position tracking. Microsoft HoloLens proved to be suitable for all AR tasks where the environment is seen, nevertheless they are very expensive and hardly to get, and 5G is supported only via a dongle that is not available in Europe or via an external router.

Another aspect that turns out to be essential is the support from hardware vendors, internet sites or the network operator. Especially with 5G, it turned out that intensive support by the telecom operator is essential for proper functionality. This is significantly different to all technologies that operate in the industrial, scientific, medical (ISM) bands, where the end users are allowed to use their systems as long as they do not exceed the limits given by the specification.

While different performance studies on multiple locations show that latency in 5G networks can be reduced to 1 ms in each direction between mobile device and edge [5], most telecom operators promise 5 ms latency, which already adds up to 10 ms if data has to be transmitted to some other device and back. At CAMPUS 02 UAS an average of 8 ms in one direction is achieved, there still seems to be potential for a speedup. What is also important is the latency between the 5G network and the remote servers. This is the main bottleneck with the current CAMPUS 02 installation since there is no edge computing integration now.

For the current three use cases the 5G network offers a much better coverage of roughly 95% compared to 70% with WLAN. Also the lower latency is an enabler especially for remote rendering, where time differences above 20 ms are perceived as too slow and only are acceptable while standing still. Also WLAN gets slower with an increasing number of devices connected, also the number of available frequency bands is exhausted.

4.1 Hindering Factors

While human remote controlling can be achieved already, there is one main factor that hinders the wide spreading of computer assisted remote controlling and rendering: Latency of the overall chain is too high, currently every component adds a significant delay that leads to a slow reaction. While the 5G connection still shows room for latency improvement, the main issue is the connection to the cloud server farm. The rendering and computation itself can be optimized by adding more nodes and/or reduction of the screen resolution or the number of items shown.

A second issue is that currently there are no embedded modules available for 5G networks, and generally, there is only a handful of 5G routers that can be used to connect TCP/IP-based devices to 5G networks. With the exception of smartphones with tethering, the other devices are designed to be home or office routers for static connections to 5G networks for fast internet access. This also influences the usability of AR glasses, since there is no product available with native 5G access.

The conducted experiments also showed that neither WLAN nor 5G are absolutely free of interruptions. For implementation of such networks one should provide means of recovery. Unmanned Ground Vehicles (UGVs) should be able to steer back to the last position with proper reception, manufacturing processes should allow intermediate stops without any harm to equipment or even persons, while for all business processes a kind of information fusion from different sources can be used to deal with interruptions of data transmission.

The last aspect that is to be mentioned is the current readiness level of private 5G net installations. While 5G in terms of being a high-speed internet connection is already available in public and a subscription is possible like with LTE, this does not make use of other potential services. To set up a campus installation involves a lot of technical understanding, resources, and willingness of the telecom provider. For a successful 5G implementation, it needs three parties: The end users, the 5G provider and the IT infrastructure and/or cloud provider.

5 Conclusion and Outlook

Real-time remote controlling over 5G will enable human operators to control multiple systems at a time from a central location, leading to shorter reaction times, less personal travel and overall process optimization. It also allows a fluid transformation from human control to AI control, where systems and vehicles are operated from cloud-based algorithms and humans only take over in case of problems.

Real-time remote rendering also shows a great potential as an application in modern factories, from additional possibilities with the above-mentioned remote controlling to break-in of new workers and especially to assist all types of personnel with whole-site information stored on a server farm, there reducing the need for high-performance devices at the point of use.

A solution for the latency issue will be the usage of edge computing, where the computation takes places directly at the wireless endpoint. While this is proposed already, most telecom operators are not able to provide the necessary experience and capabilities to allow a certain quality of service. Further research is to be conducted at the CAMPUS

02 UAS test site, moving development and experiments one step down from application layer to transportation layer, directly optimizing the 5G transmission components. The close collaboration with Citycom Graz as the network provider and Nokia as the component vendor will be a key success factor.

References

1. Smartfactory Magna Steyr. https://www.youtube.com/watch?v=F94LWRMfbG0. Accessed 01 Mar 2022
2. Mobile Knowledge UWB Kit. https://www.themobileknowledge.com/product/mk-uwb-kit-rtls/. Accessed 10 Mar 2022
3. Shi, S., Hsu, C.: A survey of interactive remote rendering systems. ACM Comput. Surv. **47**, 1–29 (2015)
4. Kelkkanen, V., Fiedler, M., Lindero, D.: Synchronous remote rendering for VR. Int. J. Comput. Games Technol. (2021)
5. Ansari, J., et al.: Performance of 5G trials for industrial automation. Electronics **2022**, 11 MDPI (2022)

Concept of a 5G Hybrid Wireless Campus Network as Testbed for Industrial Applications

Christoph Cammin(✉), Thomas Doebbert, Bettina Solzbacher, and Gerd Scholl

Electrical Measurement Engineering, Helmut-Schmidt-University, University of the Federal Armed Forces Hamburg, 22043 Hamburg, Germany
christoph.cammin@hsu-hh.de
https://www.hsu-hh.de/emt/

Abstract. The 5th generation technology standard for broadband cellular networks (5G) is currently being deployed at a large scale. In addition to the expansion of public 5G networks, (private) 5G campus networks are also set up in many areas. However, in parallel with the development of 5G, numerous other wireless solutions have also evolved, primarily using unlicensed frequency bands. In addition to various versions of the consumer grade IEEE 802.11-based WLAN standards and Bluetooth, a number of specific wireless solutions for predominantly industrial data communication have been established. The advantages of these comparatively simpler, non-5G wireless technologies are lower price levels, higher availability of established products on the market, and improved energy efficiency, which often lead to a significant commercial success and justify their beneficial use. This contribution presents a hybrid campus wireless network infrastructure, which is intended to fulfill various requirements as a testbed, especially in the areas of automation, logistics and traffic. A focus is on wireless coexistence, functional safety requirements and functionalities in conjunction with the associated security for safety.

Keywords: 5G · Wireless network · Hybrid network · IO-Link Wireless · Functional safety · Security · Coexistence

1 Introduction

In many areas of the industrial value chain, the need for secure, reliable and fast wireless communication solutions is growing. Compared to previous wired communication solutions, advantages arise in particular from greater local flexibility, the elimination of expensive and mechanically sensitive data lines, and a reduction in installation effort, which can ultimately also favor cost savings (e.g., [37]). Thereby, the 5th generation technology standard for broadband cellular networks (5G) is being marketed as a universal wireless communication solution and is currently deployed at a large scale. In addition to the expansion of public 5G networks, private 5G campus networks are also installed in many

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 458–469, 2023.
https://doi.org/10.1007/978-3-031-16281-7_43

areas. However, in parallel with the development of 5G, numerous other wireless solutions found their specific application domains, primarily using unlicensed frequency spectrum. In addition to the variants and various versions of the consumer standards such as IEEE 802.11-based WLAN (in the following referred to as WLAN), Bluetooth, and in some cases, Zigbee, a number of specific wireless standards for predominantly industrial (sensor/actuator) data communication have also become established (e.g., [15,29,37]). Within the smart traffic domain, WLAN-based 802.11p devices compete with those based on C-V2X. The advantages of these comparatively simpler, non-5G wireless technologies have usually been lower cost, higher availability of established products on the market, and improved energy efficiency. From the user's perspective, the fundamental question is: Which wireless standards and systems can provide the required quality of service (QOS)? A starting point for the testbed shall be the Digital Sensor-2-Cloud Campus Platform (DS2CCP) with an implementation of IO-Link Wireless and 5G technology for safety-related and security-for-safety applications [14,15]. Within the dtec.bw DS2CCP project future-oriented and cross-platform solutions will be developed to enable reliable and robust wireless communication of sensors and actuators from the industrial shop floor to the edge cloud, while taking into account requirements regarding safety, security and timing. The aim of the project is a consistent and holistic conception of an industrial DS2CCP for applications in the field of automation technology [21].

This paper is organized as follows: An introduction into different use cases with specific requirements is depicted in Sect. 2. An overview of wireless technologies is given in Sect. 3. In Sect. 4, major challenges for real-time wireless industrial applications, like safety, security and coexistence are outlined. A solution approach resulting in a 5G hybrid campus network as a testbed is shown in Sect. 5. Finally, a conclusion and an outlook are presented in Sect. 6.

2 Use Cases

The requirements on wireless communication systems for industrial applications are clearly classified according to their specific application domains [25,30,43], which are reviewed in the following section. Here, the focus is on factory automation, logistics and traffic, as these selections are related to further topic-specific research projects in a broader context of the DS2CCP project.

2.1 Factory Automation

Factory automation covers a wide range of use cases. However, a representative, typical scenario may be characterized by the following. One key requirement is to provide a wireless link for up to 100 (typical indoor) sensors or actuators within ranges of 10 m, corresponding to an individual manufacturing cell, in which for example a robot arm performs a certain assembly task [25,43]. The speeds are comparatively low, typically less than 10 m/s. In addition, a latency time in the order of 10 ms or below [43] with a remaining failure probability in the order of

10^{-9} or less shall be ensured. However, the exchange of a few bytes of data is often adequate, so that the net data rates in the range of a few kbit/s per radio node are sufficient.

2.2 Logistics

Logistics with cross-domain flows of goods, information and people has a manifold of different challenges in transporting, handling, storing, picking, sorting, packing and distributing and sometimes even includes traffic applications. Within the subdomain of intralogistics, a controllable or private environment, like an (indoor) factory hall or an (outdoor) plant ground can typically be assumed. Mobile applications, like automated guided vehicles (AGVs) and upcoming unmanned aerial vehicles (UAVs) are also under consideration. For these applications, localization features with an accuracy in the order of meters to centimeters are of high importance. The velocities of the mobile nodes may be up to the order of 20 m/s. Furthermore, tracking of goods is another typical application, which requires cost and energy efficient localization services with an accuracy in the 10 m to decimeter range. The required communication distances are typically 30 m to 1 km, with up to 100 nodes. The requirement with respect to latency times is around 10 ms for safety relevant applications and up to around 1 min for monitoring related applications, according to [25,43].

2.3 Traffic

In smart traffic applications most subscribers are mobile and their velocities can be, even for ground-based traffic, up to 80 m/s. The required ranges are between 10 m for local applications, such as a single traffic light control, up to several kilometers for large-scale traffic control systems. Latencies range from a few milliseconds for safety-critical applications over short distances, such as collision warning to a few seconds or even minutes for traffic flow control. The data rates lie between a few bytes per hour, such as for energy-autonomous monitoring of the occupancy of a parking lot, to about 100 Mbit/s per application when several cameras transmit live images in real time.

3 Wireless Technologies

Important drivers of 5G networks as an universal communication solution are the three main application domains: EMBB (Enhanced Mobile Broadband) with high data rates and transmission capacities for data and streaming applications, URLLC (Ultra Reliable Low Latency Communications) as a key scenario for communication with high reliability and very low latency [17] and MMTC (Massive Machine Type Communication) with a low power requirement, low complexity, low data rates and a high number and density of communication devices. In addition to these generic areas, 5G networks can be configured as needed and used in parallel with different requirements and profiles through network slicing.

Commercially available systems are usually based on the currently available Release 15, which mainly supports EMBB functions [6]. Releases 16 and 17 complete the URLLC functionalities, especially for the area of factory automation with smart factory assembly and control processes, AGVs and UAVs. URLLC utilizes low-latency mini-slots, already implemented in Release 15. Additionally, integration of time sensitive communication (TSC) is under deployment [2]. Private campus networks are available with Release 15 (and also the predecessor standard Long Term Evolution (LTE)). These private networks are of particular interest for industrial applications [4,5,44]. In Release 17, further support for private networks through neutral host models is introduced, allowing access to standalone private networks with third-party credentials, including public network operators. In particular, controllability and data sovereignty are key factors [3]. Usually they are operated in an own frequency range, managed by the user himself.

In contrast to 5G technology, non-cellular technologies such as WLAN (IEEE 802.11 series) and narrowband standards (IEEE 802.15 series) utilize the license-free ISM frequency bands. WLAN systems are typically used for wireless IP-based networks and offer high transmission capacities. In addition to WLAN in the 2.45 GHz ISM band, radio systems like Bluetooth (BT)/Bluetooth low energy (BLE) (originally IEEE 802.15.1 based) or Zigbee (IEEE 802.15.4 based) are also located in this frequency band. On the other hand, such systems are capable of connecting a very high number of radio nodes and some of them implement mesh communication. To meet industrial requirements, additional standards have been developed for both WLAN and systems such as WirelessHART [20] or EnOcean, based on the IEEE 802.15.4 standard, or IO-Link Wireless (IOLW) [22,26,45] based on the IEEE 802.15.1 standard, to address specific wireless automation domains.

The IEEE 802.11p standard is specifically employed for V2X communication, i.e. traffic applications. These systems use the intelligent transportation system (ITS) frequency band in the 5.9 GHz range and enable adhoc networks to be set up to other devices in short range in order to exchange data. As competitor, C-V2X technology uses the 4G/5G networks and also allows to connect to other global services. Furthermore, the localization features in 4G and 5G based C-V2X are promising for future applications regarding safety enhancements in traffic [19].

4 Challenges for Real-Time Wireless Industrial Applications

The overall system concept shall address relevant requirements that have to be imposed on wireless technologies in order to realize a flexible testbed. In particular, safe and secure system solutions with respect to typical safety integrity level (SIL) classifications and performance levels are a key challenge. In this context, the requirements are orientated on relevant safety protocols, such as

PROFIsafe [32,34] or IO-Link Safety. Generally, the availability of wireless spectrum is a prerequisite for reliable, low-latency wireless communication. Thus, the coexistence with other wireless systems is also a challenging task [12,29,48]. The requirements with regard to functional safety and security were discussed in [13,15] and briefly outlined here for completeness.

4.1 Safety

The testbed for industrial safety applications shall support safety integrity level (SIL) for at least SIL3 according to IEC 61508 [24] or Category 4 and PL e according to ISO 13849-1 [27]. Furthermore, a safety function response time (SFRT) in the order of 10 ms shall be achievable to cover a wide range of possible applications. The possibility of combining different SIL and non-safety devices in one communication network is necessary. A restriction of the safety transmission protocol to the safety layer by supporting requirements of IEC 61784-3 shall be focused on. The testbed needs to combine a overall safety architecture for 5G and other wireless technologies or a combination of different safety architectures depending on the transport medium. According to [15,18], it is mandatory that deterministic data transmission for a safe and reliable communication including low latency times, low jitter and low packet loss rates is provided.

4.2 Security

Another challenge for real-time wireless industrial applications are security threats, which can be classified from a physical access perspective with three main types of threats [42]:

- physical access
- access in close proximity (e.g., BLE, IOLW or Near Field Communication (NFC))
- remote access (e.g., by network via an unsecured gateway).

Besides the physical security perspective, software and key provisioning security, run-time security, and foundation for security form basic prerequisites for countering these fields of attacks, as also outlined in [15]. The usage of up-to-date cryptographic algorithms for encryption, authentication, integrity and e.g. key exchange, which also needs to run on hardware with limited power consumption are also mandatory (e.g., [13,16]).

4.3 Wireless Coexistence

Frequency planning is essential to ensure a high quality of service (e.g., [1,43,48]). In many cases, the ISM bands are used to make commercial applications available to consumers at low cost, since no license fees are incurred. However, the effect of this is that these frequency bands are often excessivly used. In contrast, the frequency bands of the mobile network providers can typically only be accessed by the providers. Therefore, in most cases, interferers are avoided.

In addition to the primary use of the 5G frequency bands, they are also allocated to other radio applications, e.g. for historical reasons, as described e.g. in [23,28,40]. In the range of public 5G-bands n7 (2.6 GHz) and n78 (3.5 GHz), for example, there is a partial overlap with satellite radio applications (e.g., [35,39,41]). However, the structure of the 5G system means that the network is designed with redundancy so that other bands can then be used, generally. In most cases, transmissions are then switched by the devices in such a way that the highest possible QOS will be achieved.

Generally, coexistence management in the 2.45 GHz and 5.8 GHz ISM bands is more difficult than in the 5G network, as several competing radio standards share the available bandwidth (see e.g. [12,29,36,38,46,47]). As the frequency bands can be used generally, no direct communication between the individual standards to achieve improved communication exist. For example, it is fundamental that actuators and sensors in the closed-loop control work together optimally. Thus, interferences of the radio connections as potential in the different frequency bands are always to be evaluated as a risk, which cannot be disregarded during the planning and implementation. However, since the ranges in these frequency bands are usually small (typically 20 m radius), good coverage can be achieved for stationary installations.

5 Solution Approach

To test and measure the performance of the various wireless technologies for the different use cases, such as robotics, AGVs, UAVs or smart traffic control systems, a comprehensive testbed including test networks is planned. A visualization of the application scenarios with the corresponding radio networks is sketched in Fig. 1.

Since, as described above, various competing radio standards are advertised as being suitable for certain applications, at least the most important wireless technologies must be available as infrastructure for a metrological comparison. Furthermore, there are systematic constraints, such as typical licensing of frequency bands for cellular radio technologies, which generates advantages like reduced interference and disadvantages like additional costs. Therefore, a wireless reconfigurable, hybrid campus network infrastructure shall offer enhancements and flexibility suited dependent on the specific application. This will consist of at least the following elements, whereby the DS2CCP, whose first concept study was presented in [15], serves as the basis: Access to public 5G networks, a private 5G campus network, WLAN infrastructure (at least with the standards n, ac and p), IOLW equipped lab environments and a hybrid V2X infrastructure, i.e. utilizing both, C-V2X and 802.11p will be the components. For large-scale applications with long-range spatial dimensions, such as for traffic flow management, public 5G networks will be a favorable approach to be tested. On a medium-scale like an industrial factory plant, private campus networks and WLAN based infrastructure is feasible with reasonable effort. Intralogistic AGVs and UAVs may also be supported by these networks as enabler technologies, especially in order

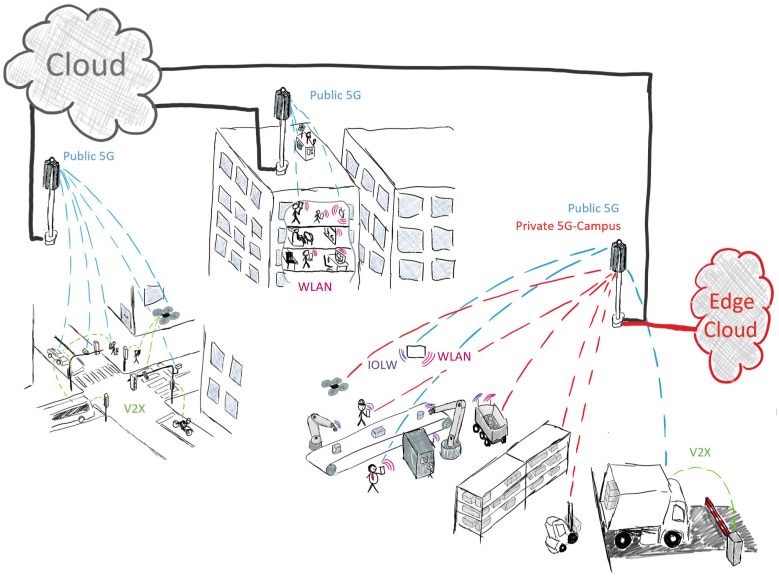


Fig. 1. Sketch of the suggested wireless scenario with solution approach.

to fulfill functional safety requirements. In the short-range, like traffic intersections or the machine-area of e.g. a robotic arm in a manufacturing environment, V2X communication and IOLW, respectively, can furthermore enhance both, safety-critical and non-safety communication.

In Fig. 2, an estimation of pure 5G technologies performance indicators within the application areas is depicted. For comparison, an estimation of performance indicators for 5G technology enhanced by corresponding non-cellular technologies is depicted in Fig. 3. The classification is made in the meta categories cost efficiency, latency, data rate, mobility, node density, energy efficiency, coverage area and managing effort, whereby markings far outside are advantageous in each case. Meta categories depict sub-aspects being grouped. For example, cost efficiency includes the costs of acquiring infrastructure, nodes, installation, operation, radio approvals/licenses, etc. In this context, latency includes the sub-figure of merits describing minimal, typical, maximal (ensured) latency and according jitter. In this context, also reliability is covered by latency. Data rate includes minimal, typical, maximal (ensured) data rate as well as throughput, etc. The respective classifications are not necessarily absolute, but should be regarded as a corresponding, expectable order of magnitude, since slightly different ranges can be achieved in each case by adjusting the configuration in the individual categories.

For enhanced 5G networks, some network components and nodes can be operated using the non-cellular technologies. Thus, the cost efficiency can significantly be increased and further advantages can be achieved mainly in the areas of node density and managing effort in relation to the nodes. In the latter,

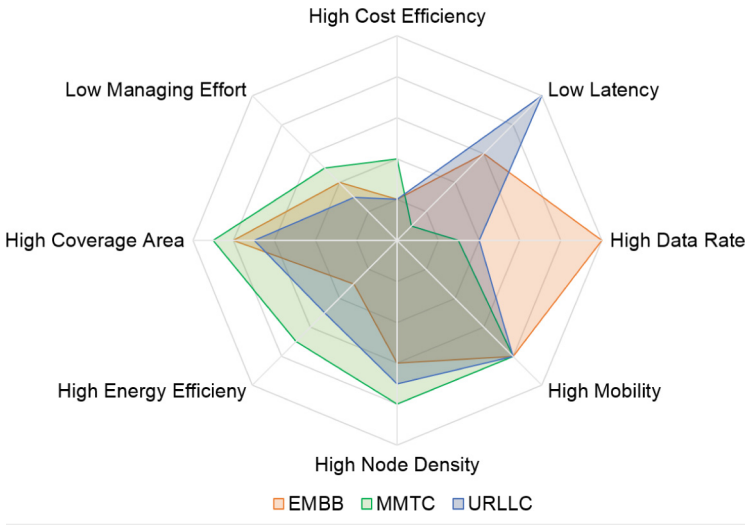


Fig. 2. Capabilities of pure 5G technology. The graphic is based on the data in [4,30, 31].

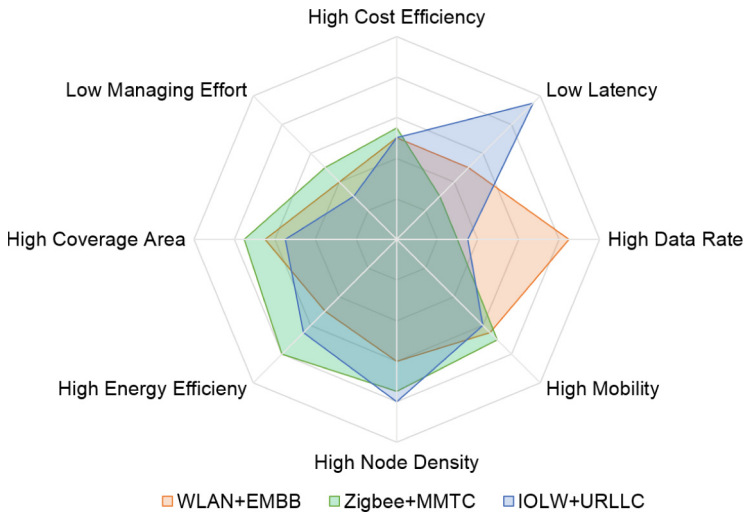


Fig. 3. Capabilities of 5G enhanced by non-cellular technologies. The graphic is based on the data in [4,30,31].

this advantage can be achieved by eliminating 5G nodes and connecting them through easier-to-manage technologies. However, the performance in terms of the meta categories latency, data rate and mobility may slightly be decreased due to transitions (e.g., in gateways) between the systems. There are significant

advantages mainly in the areas of cost efficiency, node density and managing effort in relation to the nodes.

In the focus of wireless coexistence, the individual wireless subscribers are embedded in a coexistence scenario i.e. a signal diversity. An approach to scale this down in the transfer to test components or tests of individual subscribers was shown in [7]. Furthermore a (customized to the radio standard) consideration of the corresponding radio channels as well as OTA testing is useful [8–11]. In context of a holistic security approach being partly addressed in Sect. 4, a security-by-design approach shall be implemented even though a combination of wireless technologies may induce a mapping of parameters at a lower protocol layer. OPC-UA defines a security model to describe aspects of user-authentication and controlling access rights, application identification using digital certificates, auditing user and system activities, availability and secure/encrypted message transmission [33]. Different topologies depending on client/server or publish/subscribe are application- and resource-dependent possible.

6 Conclusion and Outlook

In this paper, possible radio-based communication solutions were presented based on various application scenarios. Also considering the factors (cyber-) security, functional safety and wireless coexistence, a hybrid wireless campus network infrastructure was presented. Therefore, advantages are highlighted how a 5G technology can be enhanced/combined with a non-cellular technology. This infrastructure is currently deployed and will be evaluated as a wireless communication testbed for applications such as AGVs, UAVs, robotics or connected traffic systems to enhance safety-related and security-for-safety-related features and to improve traffic flow.

Acknowledgment. This work is an extended continuation of the work presented in a very first stage as a work-in-progress under [15]. The authors like to acknowledge R. Heynicke and H. Beuster from the Helmut Schmidt University as well as B. Kärcher from FESTO for their continuous support. Funding was granted by the Federal Ministry for Economic Affairs and Climate Action (BMWK) (formerly Federal Ministry for Economic Affairs and Energy (BMWi)) and the Project Management Jülich (PTJ) under the Wissens- und Technologietransfer durch Patente und Normen (WIPANO) program as IO-Link Wireless Standardization for IEC-Approval (IOLW-4-IEC) with grant no. 03TN0005A and by the Federal Ministry of Defense under the dtec.bw program as Digital Sensor-2-Cloud Campus Platform (DS2CCP) (project website: [21]).

References

1. American National Standard for Evaluation of Wireless Coexistence. ANSI/USEMCSC C63.27-2021 (Revision of ANSI C63.27-2017), pp. 1–87 (2022). https://doi.org/10.1109/IEEESTD.2022.9777636
2. 5G Alliance for Connected Industries and Automation (5G-ACIA). Integration of 5G with Time-Sensitive Networking for Industrial Communications (2020)

3. 5G Alliance for Connected Industries and Automation (5G-ACIA). 5G Non-Public Networks for Industrial Scenarios (2019)
4. Aijaz, A.: Private 5G: the future of industrial wireless. IEEE Indust. Electron. Magaz. **14**(4), 136–145 (2020). https://doi.org/10.1109/MIE.2020.3004975
5. Alabbasi, A., Dudda, T., Zou, Z., Kronander, J.: 5G toolbox for realizing industrial automation. In: 2019 IEEE 2nd 5G World Forum (5GWF), pp. 512–515 (2019). https://doi.org/10.1109/5GWF.2019.8911658
6. Baker, M., Poikselkä, M.: 5G Releases 16 and 17 in 3GPP. Tech. rep., Nokia Bell Labs CTO, Standardization and Research, 20 April 2020
7. Cammin, C., Krush, D., Heynicke, R., Scholl, G.: Test method for narrowband F/TDMA-based wireless sensor/actuator networks including radio channel emulation in severe multipath environments. J. Sens. Sens. Syst. **7**(1), 183–192 (2018). https://doi.org/10.5194/jsss-7-183-2018
8. Cammin, C., Krush, D., Heynicke, R., Scholl, G.: Employing correlation for wireless components and device characterization in reverberation chambers. J. Sens. Sens. Syst. **8**(1), 185–194 (2019). https://doi.org/10.5194/jsss-8-185-2019
9. Cammin, C., Krush, D., Heynicke, R., Scholl, G.: Reproducibility of fading propability in a reverberation chamber for wireless device testing. In: 2019 IEEE Radio and Antenna Days of the Indian Ocean (RADIO), pp. 1–2 (2019). https://doi.org/10.23919/RADIO46463.2019.8968873
10. Cammin, C., Krush, D., Heynicke, R., Scholl, G.: Deep fading in a reverberation chamber for wireless device testing. IOP Conf. Ser. Mater. Sci. Eng. **766**(1), 012004 (2020). https://doi.org/10.1088/1757-899x/766/1/012004
11. Cammin, C., Krush, D., Heynicke, R., Scholl, G.: Sensing reverberation chamber loading for IO-Link Wireless testing. In: 2021 International Conference on Electromagnetics in Advanced Applications (ICEAA), pp. 87–91 (2021). https://doi.org/10.1109/ICEAA52647.2021.9539761
12. Cammin, C., et al.: Coexisting wireless sensor networks in cyber-physical production systems. In: 2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1–4 (2016). https://doi.org/10.1109/ETFA.2016.7733593
13. Doebbert, T.R., Cammin, C., Scholl, G.: Precision measurement of the application-dependent current consumption of a wireless transceiver chip in the time and frequency domain. J. Sens. Sens. Syst. **11**(1), 149–159 (2022). https://doi.org/10.5194/jsss-11-149-2022
14. Doebbert, T.R., Cammin, C., Scholl, G.: Safety architecture proposal for low-latency sensor/actuator networks using IO-Link Wireless. IEEE Access **10**, 3030–3044 (2022). https://doi.org/10.1109/ACCESS.2021.3128758
15. Doebbert, T.R., Cammin, C., Scholl, G., Kärcher, B.: Study of a safe and secure ecosystem based on IO-Link Wireless and a 5G campus network. In: 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1–4 (2021). https://doi.org/10.1109/ETFA45728.2021.9613484
16. Doebbert, T.R., Krush, D., Cammin, C., Jockram, J., Heynicke, R., Scholl, G.: IO-Link Wireless device cryptographic performance and energy efficiency. In: 2021 22nd IEEE International Conference on Industrial Technology (ICIT), vol. 1, pp. 1106–1112 (2021). https://doi.org/10.1109/ICIT46573.2021.9453590
17. Esswie, A., Pedersen, K.: Analysis of Outage Latency and Throughput Performance in Industrial Factory 5G TDD Deployments (2020). arXiv preprint arXiv:2012.05507

18. Etz, D., Frühwirth, T., Ismail, A., Kastner, W.: Simplifying functional safety communication in modular, heterogeneous production lines. In: 2018 14th IEEE International Workshop on Factory Communication Systems (WFCS), pp. 1–4 (2018). https://doi.org/10.1109/WFCS.2018.8402371
19. Halili, R., Weyn, M., Berkvens, R.: Comparing localization performance of IEEE 802.11p and LTE-V V2I communications. Sensors **21**(6) (2021). https://doi.org/10.3390/s21062031
20. HART-FieldComm Group. WirelessHART (2021)
21. Helmut-Schmidt-University: Digital Sensor-2-Cloud Campus Platform (2022). Project website: https://dtecbw.de/home/forschung/hsu/projekt-ds2ccp/projekt-ds2ccp. Accessed 18 June 2022
22. Heynicke, R., et al.: IO-Link Wireless enhanced factory automation communication for Industry 4.0 applications. J. Sens. Sens. Syst. **7**(1), 131–142 (2018). https://doi.org/10.5194/jsss-7-131-2018
23. Hong, H., Choi, S.W., Sup Kim, C., Chong, Y.J.: Interference measurement between 3.5 GHz 5G system and radar. In: 2018 International Conference on Information and Communication Technology Convergence (ICTC), pp. 1539–1541 (2018). https://doi.org/10.1109/ICTC.2018.8539422
24. IEC 61508: Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems (E/E/PE, or E/E/PES)
25. IEC 62657-1:2017. Industrial Communication Networks - Wireless Communication Networks -Part 1: Wireless Communication Requirements and Spectrum Considerations (2017). https://webstore.iec.ch/publication/33125. Accessed 16 June 2021
26. IO-Link Community. IO-Link Wireless System Extensions - Specification Draft V1.1.3 for Review, December 2021, Order No: 10.112 (2021). https://io-link.com/share/Downloads/System-Extensions/IO-Link_Wireless_System_Specification_10112_d113_Dec21.pdf. Accessed 06 June 2022
27. ISO 13849–1:2015. Safety of machinery - Safety-related parts of control systems - Part 1: General principles for design
28. Kim, S., Visotsky, E., Moorut, P., Bechta, K., Ghosh, A., Dietrich, C.: Coexistence of 5G with the incumbents in the 28 and 70 GHz bands. IEEE J. Select. Areas Commun. **35**(6), 1254–1268 (2017). https://doi.org/10.1109/JSAC.2017.2687238
29. Krush, D., Cammin, C., Doebbert, T.R., Heynicke, R., Scholl, G.: Coexistence management methods and tools for IO-Link Wireless. In: 2021 17th IEEE International Conference on Factory Communication Systems (WFCS), pp. 151–158 (2021). https://doi.org/10.1109/WFCS46889.2021.9483594
30. Krush, D., Cammin, C., Heynicke, R., Scholl, G., Kaercher, B.: A wireless communication system for energy and environmental monitoring. J. Sens. Sens. Syst. **6**(1), 19–26 (2017). https://doi.org/10.5194/jsss-6-19-2017
31. Nixon, M.: A Comparison of WirelessHART and ISA100.11a (2012). http://www2.emersonprocess.com/siteadmincenter/PM%20Central%20Web%20Documents/wirelesshart-vs-isa-WP.pdf. Accessed 06 June 2022
32. OPC Foundation. OPC UA Online Reference
33. OPC Foundation. OPC Unified Architecture Part 2: Security Model, Release 1.04
34. Profinet University. PROFIsafe Functional Safety
35. Rohde and Schwarz. Coexistence of 5G and satellite services in the C band (2019)
36. Sakal, I., Simunic, D.: Simulation of interference between Bluetooth and 802.11b systems. In: 2003 IEEE International Symposium on Electromagnetic Compatibility (EMC 2003), 2003, vol. 2, pp. 1321–1324 (2003). https://doi.org/10.1109/ICSMC2.2003.1429164

37. Scholl, G., Heynicke, R., Krueger, D., Hornung, R.: Wireless automation. In: Proceedings SENSOR 2013, pp. 379–383 (2013). https://doi.org/10.5162/sensor2013/C3.1
38. Solzbacher, T., Heynicke, R., Scholl, G.: Parallel processing of RSSI signals for gapless monitoring of the 2.45 GHz ISM band. tm - Technisches Messen 85(s1), s124–s128 (2018). https://doi.org/10.1515/teme-2018-0047
39. Son, H., Chong, Y.: Coexistence of 5G system with Fixed satellite service Earth station in the 3.8GHz Band. In: 2018 International Conference on Information and Communication Technology Convergence (ICTC), pp. 1070–1073 (2018). https://doi.org/10.1109/ICTC.2018.8539462
40. Son, H., Chong, Y.: Analysis of the interference effects of 5G system on automotive collision avoidance radars. In: 2019 International Conference on Information and Communication Technology Convergence (ICTC), pp. 1463–1466 (2019). https://doi.org/10.1109/ICTC46691.2019.8939683
41. Tan, H., Liu, Y., Feng, Z., Zhang, Q.: Coexistence analysis between 5G system and fixed-satellite service in 3400–3600 MHz. China Commun. **15**(11), 25–32 (2018). https://doi.org/10.1109/CC.2018.8543046
42. Texas Instruments. Building your application with security in mind (2020). https://www.ti.com/lit/ml/swpb020e/swpb020e.pdf?ts=1611941696435&. Application Note. Accessed 02 Feb 2022
43. VDI/VDE 2185. Radio based communication in industrial automation. (all parts)
44. Vitturi, S., Zunino, C., Sauter, T.: Industrial communication systems and their future challenges: next-generation ethernet, IIoT, and 5G. Proc. IEEE **107**(6), 944–961 (2019). https://doi.org/10.1109/JPROC.2019.2913443
45. Wolberg, D., Rentschler, M., Gaggero, P.: Simulative performance analysis of IO-Link Wireless. In: 2018 14th IEEE International Workshop on Factory Communication Systems (WFCS), pp. 1–10 (2018). https://doi.org/10.1109/WFCS.2018.8402352
46. Zhang, H., Chu, X., Guo, W., Wang, S.: Coexistence of Wi-Fi and heterogeneous small cell networks sharing unlicensed spectrum. IEEE Commun. Magaz. **53**(3), 158–164 (2015). https://doi.org/10.1109/MCOM.2015.7060498
47. Zhang, R., Wang, M., Cai, L.X., Zheng, Z., Shen, X., Xie, L.: LTE-unlicensed: the future of spectrum aggregation for cellular networks. IEEE Wirel. Commun. **22**(3), 150–159 (2015). https://doi.org/10.1109/MWC.2015.7143339
48. ZVEI. German Electrical and Electronic Manufacturers' Association: Coexistence of Wireless Systems in Automation Technology (2009)

Reference Network and Localization Architecture for Smart Manufacturing Based on 5G

Stephan Ludwig[1(✉)], Doris Aschenbrenner[1,2], Marvin Scharle[1,3], Henrik Klessig[4], Michael Karrenbauer[5], Huanzhuo Wu[6], Maroua Taghouti[7], Pedro Lozano[8], Hans D. Schotten[5], and Frank H. P. Fitzek[6]

[1] Aalen University, 73430 Aalen, Germany
Stephan.Ludwig@hs-aalen.de
[2] TU Delft, Faculty of Industrial Design Engineering, 2628CE Delft, The Netherlands
[3] Conclurer GmbH, 89522 Heidenheim an der Brenz, Germany
[4] Robert Bosch GmbH – Corporate Research, 71272 Renningen, Germany
[5] Technische Universität Kaiserslautern, 67663 Kaiserslautern, Germany
[6] Technische Universität Dresden, 01062 Dresden, Germany
[7] Technische Universität Berlin, 10623 Berlin, Germany
[8] Ericsson GmbH, 52134 Herzogenrath, Germany

Abstract. 5G promises to shift Industry 4.0 to the next level by allowing flexible production. However, many communication standards are used throughout a production site, which will stay so in the foreseeable future. Furthermore, localization of assets will be equally valuable in order to get to a higher level of automation. This paper proposes a reference architecture for a convergent localization and communication network for smart manufacturing that combines 5G with other existing technologies and focuses on high-mix, low-volume applications, particularly at small and medium-sized enterprises. The architecture is derived from a set of functional requirements, and we describe different views on this architecture to show how the requirements can be fulfilled. It connects private and public mobile networks with local networking technologies to achieve a flexible setup addressing many industrial use cases.

Keywords: Adaptive manufacturing · 5G · Localization · M2M

1 Introduction

Industry 4.0 or the Industrial Internet of Things (IIoT), will merge physical and real worlds as connected cyber-physical production systems (CPPSs) that will enable an unprecedented production efficiency and degree of automation, as well as highly flexible and reconfigurable "smart" manufacturing—including quality management. In order to stay competitive in global markets, companies have to be flexible and able to change production easily. Especially small and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 470–479, 2023.
https://doi.org/10.1007/978-3-031-16281-7_44

medium-sized enterprises (SMEs) supply large enterprises with customized parts as intermediate goods, and they produce in a high mix, low volume (HMLV) production at small batches. In that, they must efficiently retool and optimize set-up times, as around 1.8 h of manufacturing time are currently lost on average at German production plants of the electronics and metal processing industry [8]. Additionally, precise real-time localization of goods is foreseen to significantly cost-optimize intralogistics. The 5th generation of mobile communications (5G) will be a key factor therein. While an architecture for 5G has been developed by 3GPP, smart manufacturing use cases have been proposed in the 5G ACIA [3]. Subsequent requirements are fulfilled in 3GPP Rel-16 onwards (available now) [1]. Rel-18 (finished by mid of 2024) should allow for localization with precision below 1 m [2].

Introducing 5G, which includes private networks on premise in dedicated spectrum, faces hurdles and challenges, which often relate to *How can a 5G network be to (securely) integrated with a (existing) manufacturing environment?* In this, the literature contains reference architectures: With the Industrial Internet Reference Architecture (IIRA) [12] production technologies vertically integrate with the IIoT. Its different views respect different dimensions: system characteristics, Functional domains and crosscutting functions, in particular, connectivity. The Reference Architectural Model of Industrie 4.0 (RAMI 4.0) [6] is complementary to the IIRA and addresses dimensions at higher-level layers, e. g. the life cycle value stream, communications and hierarchy levels, which essentially resemble the automation pyramid. However, existing architectures define only few regarding the integration of operation technology (OT) and information and communication technology (ICT) like 5G into an industrial network, for which interdisciplinary joint work from different fields like manufacturing, automation, manufacturing IT, IT security and communication technology is required. Although there are already promising approaches to use 5G to enable more flexible (and more dynamic) manufacturing processes [7], mainly large companies carry out these examples and target large production lines. Nevertheless, there are still huge differences between SMEs and large enterprises [15,20], specifically regarding new technological approaches.

This is why we suggest providing a 5G production "small cell" with localization and context-awareness. In comparison to existing 5G applications, our approach uses flexible small cells without full area coverage, but only where needed, that enable mutual federation. In Sect. 2, we help create a mutual understanding by summarizing interdisciplinary requirements, especially from a plant's manager's perspective. Based on [11], we propose a reference network and localization architecture in Sect. 3 with 5G as its nucleus and specifically targeting HMLV needs, which fulfills the requirements and integrates well into the IIRA and RAMI 4.0 (cf. [11]). The 5G small cell evolves into a heterogeneous industrial localization and communication network, expanding the components-based architecture in [11,13] by virtualization, data flow, security (leveraging strong 5G security features [4]) and management views. With this dynamic constellation are enabled on the shopfloor, representing the reality of SMEs with HMLV manufacturing requirements. Section 4 summarizes the article and outlines further research.

2 Summary of Functional System Requirements

When plant managers consider a new system, they weigh out the economic benefits with monetary effort (return on invest), while assessing how technical requirements are fulfilled. In most cases, there is no *one and only* killer use case that pays off the invest. Furthermore, a smooth transition from legacy solutions to new technology is necessary, requiring interoperability across systems. We focus on related requirements in this section originating from in-field experience collected by the 5GANG project [5].

Req-A) Interoperability and Network/Localization Convergence: A new system shall support use cases with a common and interoperable infrastructure, ideally deployable worldwide—while it is understood that user equipment (UE) implements only subsets of the standard and hence supports not all technical requirements of all use cases simultaneously. Especially SME hesitate to spend the significant costs for a full coverage 5G network on their production premises. Flexible small cells could be used just there, where connectivity is required, and multiple small cells shall act as a federated system with non-disruptive handover. Since plant managers have a plethora of existing networking technologies already, another requirement relates to network convergence. The system shall transparently integrate new devices without adjustments for incremental retrofitting, which also applies for different existing wired or wireless access technologies like Industrial and classic Ethernet, field bus networks, WiFi, Bluetooth and low-power wide area network (LPWAN) (LoRa, mioty etc.). Although 5G allows end devices to specialize into application specific need using enhanced mobile broadband (eMBB), ultra-reliable and low latency communications (uRLLC) and massive machine type communications (mMTC), while a base station provides all three flavors, applications can have more strict requirements than 5G might fulfill, e. g. in battery lifetime, where (ultra) LPWAN technologies like LoRa or mioty might outperform 5G mMTC. With this and gateway support, devices establish multiple connections to different base stations for load balancing, seamless handover, extended coverage, mesh networks, 5G device-to-device communication and increased reliability. Besides communication, localization support of assets enables an optimization of the intralogistics. Hence, future systems shall converge with communications and shall support multiple localization technologies. In addition, this 5G++ flexible small cell ("5G++ FlexiCell") shall integrate into the production network and optimize the end-to-end (E2E) data transfer performance.

Req-B) Flexible Deployment and Operator Models: Plant managers require flexible network deployment, ownership and operator models, which facilitate easy and fast deployment, operation and management. This needs to be ensured in SMEs, which do not afford owning network infrastructure, as well as in large-size companies, which might want to operate their own network. Independent of the operator model, sensitive production data (which is the key value and know-how of production) shall stay on-premise, reducing the risk of data breach. To

ensure global applicability and flexibility, private networks shall be supported in licensed and unlicensed bands and in mobile network operator (MNO)-owned or dedicated, local spectrum. Finally, indoor and outdoor operation needs to be supported, e. g. for inbound logistics use cases.

Req-C) Manageability: Infrastructure personnel might not always be available in SMEs, or may be centralized in large companies. Therefore, automatic device integration and on-boarding and remote device management are key requirements. In order to allow for supervisory control of the entire system, the CPPS shall incorporate humans (cognitive strength) together with machines (procedural strength) [19]. In order to hierarchically manage them along their lifecycle, the configuration of the system components must be understood. Exposing management interfaces is another requirement enabling the interaction of industrial facilities and 5G networks. In CPPSs, the large amount of available data shall lead to automation components adopting themselves. This can be achieved by location-based services e. g. using the specific context.

Req-D) Extensibility: A flexible production requires a flexible, extensible and future-proof network technology. When services are incrementally added on the factory floor, easy and fast extension of network functions, processing resources and functionalities is required. Here, different hierarchies of network, compute and storage resources with interoperable stacks prevent congestion and component redundancy increases reliability. The dynamic environments of a flexible production necessitate efficient and flexible ways of smart and autonomous data routing, as well as automatic (edge cloud) application relocation. Related to the latter, the system shall provide a lifecycle management of applications and functions to guarantee quality of service (QoS) for different edge services sharing the same physical resources.

Req-E) Cyber Security: End-to-end confidentiality and integrity protection mechanisms are key for plant managers, as production and process information belong to their major assets. Likewise important is the prevention of fake gNB man-in-the-middle attacks and of mechanisms to purposefully disable certain services, e. g. firewalls. In addition, strong mechanisms for authentication and authorization of network entities, end devices and their rights have to be provided, which is also associated with a definite identification of subscribers. Finally, mechanisms for surveillance and monitoring of sensors, creation of log files for transparent and accountable interaction and audit trail records, i. e. chronological documentation of activities that could have affected a particular operation or event during production, are necessary.

Req-F) Traffic Separation and Isolation: Manufacturing IT networks are usually segmented for performance and security reasons, in particular according to IEC 62443. This becomes an even more important factor in light of uRLLC services and converged networks and necessitates building isolated sub-networks or slices with physical and logical separation of data storage and processing resources for

applications that are not interacting with each other or have no common data base. Also, setting up inter-site connections with connectivity along logistics routes with trustworthy environments for traffic management at network borders is an important factor.

3 Reference Architecture Overview

Reference means that the reference architecture acts as a blueprint for system architectures, which are tailored to actual system requirements. Because plant managers' requirements are that diverse, the reference architecture needs to include a plethora of different views. For an overview, cf. Fig. 1, of which present an excerpt of the most important views subsequently. The component view represents the physical components with their functionalities, which are abstracted via virtualized components and functions in the virtualization view. A third perspective is the data flow view, which describes the routing of generated data from a data provider (e. g. sensors) via a data aggregation entity towards a cloud service along the hierarchy levels of the RAMI 4.0. Due to space limitations, we omit further details on the data flow view. The IT security view contains every component and functionality related to the IT security of the architecture and addresses all the other layers while going beyond the application layer. Finally, supervisory control and asset management is addressed by the management view. The presented views of the networking architecture can to a certain extent be projected to the four bottom layers of the RAMI 4.0 [6], which immediately suggests its practicability.

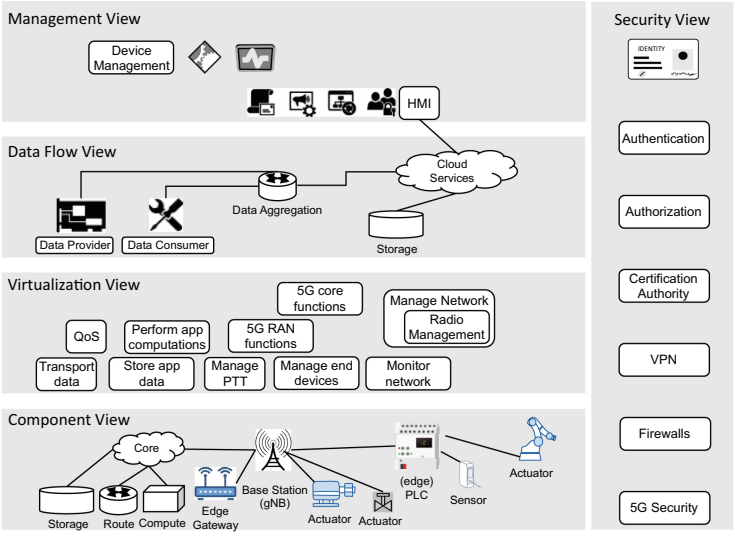


Fig. 1. Reference architecture: overview

3.1 Component View

The reference architecture includes various components and devices on multiple levels, which have been discussed in [11,13] and which address many requirements including Req-A, Req-B and Req-F for more details. Of course, each component can be present in the system multiple times, as the labels connecting them indicate. Its central builds the 5G++ FlexiCell (cf. Fig. 2), which is a 5G base stations connected to its local 5G core, such that the FlexiCell can be operated in nomad/island mode, supporting all flavors of 4G/5G (solid lines). For each production station, where 5G is currently required, a FlexiCell can be set up for local-only coverage. The 5G core via software-defined network (SDN) integrates multiple heterogeneous (dashed lines) wired communication technologies like Industrial Ethernet (e. g. Profinet or time-sensitive networking, TSN) classic Ethernet (ETH), field bus networks as well as wireless ones like WiFi, Bluetooth, mioty, LoRa and mesh networks. Data transmission can be optimized in an E2E fashion across different technologies. Sensors, Actors, (edge) programmable logic controllers (PLCs) can be connected vie 4G/5G or other technologies. 5G edge gateways offer another type of data path, particularly suited for event messaging. Furthermore, localization technologies like 5G, mioty (via radio frequency), ultrasound or RADAR are integrated into the FlexiCell. By fusion of different localization technologies position resolution can be improved, and the position data is provided through a technology-agnostic interface, which enables proximity detection and location-based services.

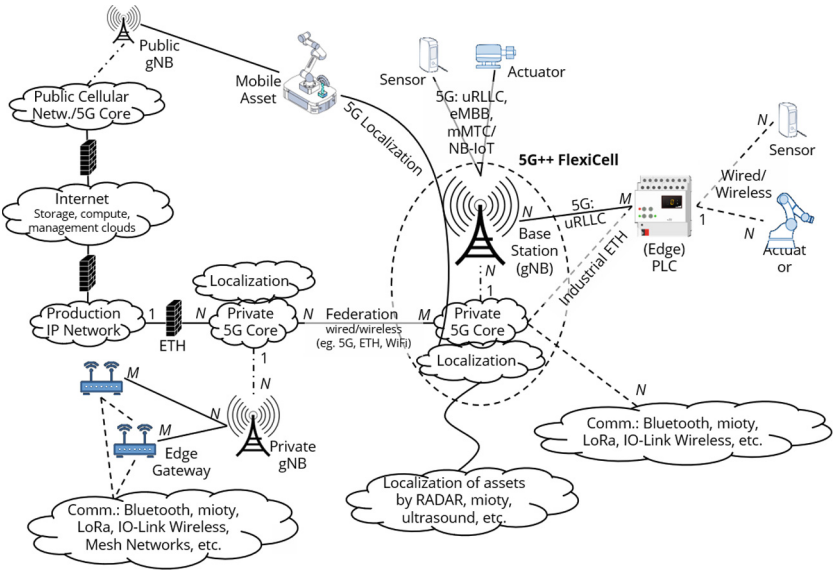


Fig. 2. Reference architecture: component view

The 5G cores of multiple 5G++ FlexiCells can federate with each other through wired or wireless connections, thus enabling flexible radio coverage with

seamless handover. Via Ethernet connections, FlexiCells can connect to the production IP network and further to the Internet and public cellular networks (all guarded by appropriate IT security measures like firewalls). Devices might (also in parallel) connect to a public 5G network, if their access rule allows this. A public–private handover (e. g. indoor–outdoor) is possible, if supported by both networks. The FlexiCell provides a radio management interface (not shown) for E2E optimization of heterogeneous data links, see Sect. 3.4

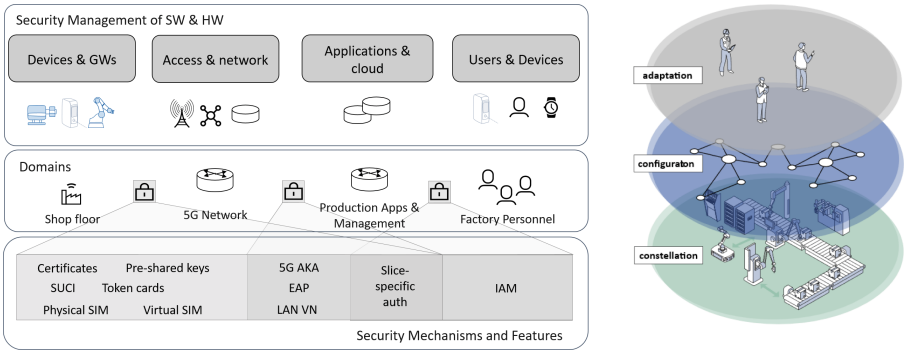
3.2 Virtualization View

Resonating with the widespread use of network function virtualization (NFV), cloud-native architectures and SDN in 5G networks [17], we follow the concept of using commercial-off-the-shelf (COTS) compute, storage and networking hardware. While benefiting from low-cost hardware, network functions and edge applications can flexibly and efficiently share the federated (cf. Sect. 3.1) hardware resources (Req-D), which could be logically and physically distributed across the factory. This is realized by a distributed cloud interconnected through different wireless and wired technologies, which enables effective balancing of resource loads across the factory. The virtualization view details how the plant-wide management of physical resources is organized and acts as an intermediate view between the components and the data flow view. Along the lines of ETSI reference architectures for NFV [9] and multi-access edge computing (MEC) [10], our architecture consists of three entities, which interact with and manage components, virtual functions and edge applications. As an important distinction from the general NFV view, we consider sensors and actuators deployed across the plant to be virtualized, as well (e. g. as digital twins). With this, data provided by massive networks of multipurpose wireless sensors can be shared among different industrial applications facilitating an efficient reuse of machines, robots, etc., and reducing costs for plant managers.

Due to space limitations, we omit further details on the Data flow view.

3.3 Security View

The 5G security architecture is designed to provide protection of connected devices and individuals' privacy (Req-E, Req-F). Our reference architecture naturally integrates 5G security mechanisms beyond traditional physical SIM cards (cf. Fig. 3a). Physical SIM cards would still continue to be used for mobile broadband subscribers with smartphones, whereas virtual SIM cards are used for uRLLC and mMTC. Non-SIM-card-based credentials are useful for integration with existing devices (Req-A) and users. In the MNO domain, two alternatives exist, namely the 5G authentication and key agreement (5G AKA) protocol and the extensible authentication protocol (EAP) framework. This strong authentication and key agreement procedure also applies for authentication, which the operator delegates to a third party (Req-E). By using these mechanisms, components (devices, users and applications) of both worlds can be securely interconnected on different hierarchy levels.



(a) Security View

(b) Management View with layers of information

Fig. 3. More reference architecture views

Internal subscriber identification can be done through (e)SIM and digital certificates, and identity and access management systems could be integrated with the 5G system through APIs. 5G also includes mechanisms to establish trusted network segments with restricted access, among which are LAN virtual networks (VN) and network slicing. Especially those address many of the security concerns, including an appropriate coverage of the IEC 62443 and private communication across sites. Furthermore, information technology (IT)/OT integration (management and application) is subject to a study in 3GPP TR 23.745.

3.4 Management View

A view on human supervisory and management functions is shown in Fig. 4, which are grouped into device, network function, edge cloud and application functions, on the one hand. On the other hand, the functions are grouped according to the lifecycle of their (virtual) components (Req-C, Req-D). For the sake of brevity, we just name some functions and refer to the figure for a longer, albeit non-exhaustive list. It is clear that not every device will support all functions, but that the management plane is able to cope with the devices' capabilities. Furthermore, we assume that the management interface supports users in efficiently solving their task with a least amount of knowledge about network technology (Req-A) such that the effort of human interaction is minimized.

The corresponding asset and configuration management is divided into three layers (cf. Fig. 3b): i) classic asset management ("constellation level") keeping locations of the devices and hardware configurations; ii) software configuration of all assets, including program code of (collaborative) robots, which allows for a container-based approach, when distribution it to the devices. iii) a newly proposed adaptation layer; here humans interact with the devices. These layers are considered being orthogonal to those of IT ecosystems, which apply here too [18]: i) operational: report key performance indicators and govern shop-floor

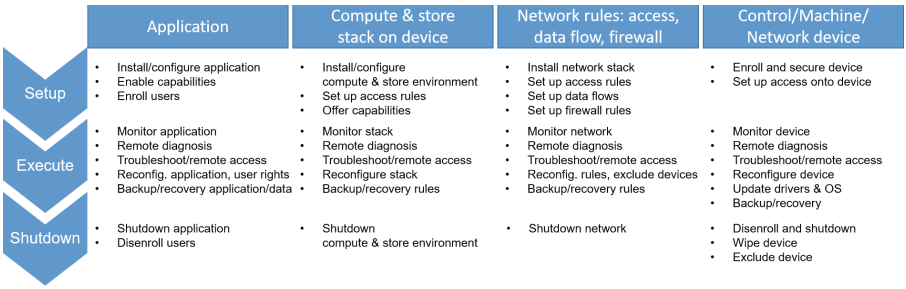


Fig. 4. Reference architecture: management view

activities; ii) tactical: transforms long-term objectives to short-term decisions; and iii) strategic level: support decisions on capital investment by the asset portfolio of the organization. Then, knowledge can be generated and managed by asset and configuration management, one can achieve a balance between costs, risks and business performance [14]. Furthermore, it enables self-X features, which keeps complex CPPS controllable [16].

4 Summary

Plant managers are still having many concerns and requirements that potentially hinder a widespread roll-out of 5G networks in factories. However, many of these requirements can be addressed, when we look at the whole network architecture, instead of considering 5G only as yet another wireless technology. In fact, we can see the 5G++ FlexiCell as a kind of glue for the convergence of different already existing communication and localization technologies. We illustrated this in the form of a reference network and localization architecture, which was in parts developed by the project "5GANG", funded by the German Federal Ministry of Education and Research, and which is currently built upon in the project "5G-FlexiCell", funded by the Federal Ministry for Economic Affairs and Climate Action. Together with company partners, the proposed architecture will be evaluated in three real-life scenarios: i) a reconfigurable cobot-machine tending in plastic molding; ii) a highspeed data sensor on an industrial manipulator in quality control; and iii) intralogistics carried out by a mobile manipulator.

References

1. 3GPP: Release 16 (2020). https://www.3gpp.org/release-16
2. 3GPP: Release 18 (2022). https://www.3gpp.org/release18
3. 5G ACIA: White Paper 5G for Connected Industries and Automation. Technical report (2018)
4. 5G Americas: The Evolution of Security in 5G. White paper (2018)

5. 5GANG Consortium: Research Project by BMBF: 5GANG – 5G for the Industry. https://www.schildknecht.ag/en/research-project-by-bmbf-german-federal-ministry-of-education-and-research-5gang-5g-for-the-industry/
6. Adolphs, P., et al.: Reference architecture model Industrie 4.0 (RAMI4.0). ZVEI and VDI, Status report (2015)
7. Cheng, J., Chen, W., Tao, F., Lin, C.L.: Industrial IoT in 5G environment towards smart manufacturing. J. Ind. Inf. Integr. **10**, 10–19 (2018)
8. Conrad, R.W.: Rüstzeitoptimierung. Technical report, Institut für angewandte Arbeitswissenschaften e. V. (2018)
9. ETSI: Network functions virtualisation (NFV): Architectural framework. Group Specification ETSI GS NFV 002 v1.2.1 (2014)
10. ETSI: Multi-access edge computing (MEC): framework and reference architecture. Group Specification ETSI GS MEC 003 v2.1.1 (2019)
11. Karrenbauer, M., et al.: Future industrial networking: from use cases to wireless technologies to a flexible system architecture. at-Automatisierungstechnik **67**(7), 526–544 (2019)
12. Lin, S.W., et al.: The Industrial Internet of Things Volume G1: Reference Architecture V1.9. Technical report, Industrial Internet Consortium (2019)
13. Ludwig, S., et al.: A 5G architecture for the factory of the future, pp. 1409–1416 (2018)
14. Maletič, D., Pačaiová, H., Nagyová, A., Maletič, M.: The link between asset risk management and maintenance performance: a study of industrial manufacturing companies. Qual. Innov. Prosper. **24**(3), 50 (2020)
15. Masoodab, T., Sonntaga, P.: Industry 4.0: adoption challenges and benefits for SMEs. Comput. Ind. **121**, 103261 (2020)
16. Müller-Schloer, C., von der Malsburg, C., Würt, R.P.: Organic computing. Informatik-Spektrum **27**(4), 332–336 (2004)
17. Ordonez-Lucena, J., Ameigeiras, P., Lopez, D., Ramos-Munoz, J.J., Lorca, J., Folgueira, J.: Network slicing for 5G with SDN/NFV: concepts, architectures, and challenges. IEEE Commun. Mag. **55**(5), 80–87 (2017)
18. Polenghi, A., Roda, I., Macchi, M., Pozzetti, A.: A conceptual model of the IT ecosystem for asset management in the global manufacturing context. In: Lalic, B., Majstorovic, V., Marjanovic, U., von Cieminski, G., Romero, D. (eds.) APMS 2020. IAICT, vol. 592, pp. 711–719. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-57997-5_82
19. Romero, D., Bernus, P., Noran, O., Stahre, J., Fast-Berglund, Å.: The Operator 4.0: human cyber-physical systems & adaptive automation towards human-automation symbiosis work systems. In: Nääs, I., et al. (eds.) APMS 2016. IAICT, vol. 488, pp. 677–686. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-51133-7_80
20. Spena, P.R., Holzner, P., Rauch, E., Vidoni, R., Matt, D.T.: Requirements for the design of flexible and changeable manufacturing and assembly systems: a SME-survey. Procedia CIRP **41**, 207–212 (2016)

Survey on Usage of 5G Campus Networks in Intralogistics

Lara Nehrke(✉), Simone Neumann, Alexandra Cieslak, and Alice Kirchheim

Helmut-Schmidt-University/University of the Federal Armed Forces, Holstenhofweg 85, 22043 Hamburg, Germany
logistiksysteme@hsu-hh.de

Abstract. In this study, the status quo of wireless communication technologies in intralogistics in Germany is investigated. What are companies' plans for the future? Which potentials and challenges are associated with them? In addition, the study examines the extent to which intralogistics companies are already using or considering 5G campus networks. To answer these questions, a qualitative survey of German intralogistics companies was conducted. It turns out that currently WLAN is most applied technology for wireless communication in intralogistics. 5G is still at an early stage. It is not very widespread, especially among small companies. None of the interviewees stated that their company uses 5G in daily operational business so far. Some companies are testing its use in pilot projects. Currently, the companies lack on quantification of the benefits of implementing 5G as well as software and hardware. According to the survey, use cases for 5G can be an alternative when existing wireless communication systems reach their limits. The findings of this qualitative study can serve as a basis for a quantitative study.

Keywords: 5G · Use cases · Intralogistics · Wireless communication

1 Introduction

Compared to other wireless communication technologies such as previous generations of mobile communications, 5G has further technological features. The implementation of this technology enables data services with higher data rates, ultra-reliable and low latency communication as well as massive machine type communication [1].

The following passage provides an overview of studies conducted in recent years in association with the implementation status of 5G in Germany in general and the potential of this technology for the intralogistics. In 2019, bitkom [2] conducted a survey on 5G and asked industrial companies in Germany about the relevance of this technology and the reasons why it is not being used. According to Franken et al. [3] 5G appears promising with regard to the digitalization and automatization of operational processes for industrial use. In an online survey by Winkler and Zinsmeister [4] on trends in the digitalization of intralogistics, experts of this field from industry and science indicated 5G as a relevant trend. In a comprehensive report by Fottner et al. [5] on the state of the art of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 480–488, 2023.
https://doi.org/10.1007/978-3-031-16281-7_45

autonomous systems in intralogistics, the capacity of wireless communication networks is determined as a challenge to be overcome in order to achieve maximum productivity. At the end of 2021, two years after the start of licensing, Franken et al. [3] asked in a study about the status quo of the expansion of 5G-campus networks in Germany. In a report based on a survey on allocation holders, it was shown that the process of implementation of 5G in industries has just begun. Here it was stated: "Only about one third of the license holders are companies integrating applications with 5G-campus networks into their operational processes. Expert interviews indicate that the demand for licenses will increase when more feasibility studies for the different sectors and more information on economic and technical aspects of campus networks are available" [3]. Automated Guided Vehicles (AGVs) are identified as the most important application to date. It also emphasizes the importance of intralogistics in the context of 5G. The project "5Guarantee" by Boshoff et al. [6] presented a study on possible applications of 5G in production environments. They compared it with other communication technologies to enable an assessment of its economic viability. For example, mobile robot platforms are being considered for intralogistics [6]. Evcenko and Kett [7] described an approach to assess the added value of 5G. That enables a 5G-benefit assessment based on use cases.

Until mid-January 2022, 186 allocations of local frequency use in the 3.7 to 3.8 GHz frequency range have been registered with the Federal Network Agency in Germany to enable companies to set up their own campus networks [8], despite the fact that applications by the Network Agency were accepted since the end of 2019 [9]. Only such a low number of frequency allocations were applied for and granted during the last two years despite the high potential of 5G especially in intralogistics. So far the scientific literature on the status quo of the implementation and application of 5G in intralogistics in Germany is scarce According to the current state of knowledge, there are no qualitative surveys on this topic. This leads to the following research questions for this study:

1. To what extent have companies in intralogistics in Germany already applied or considered 5G and to what extent will the use of this technology play a role in the future development?
2. What is the current status and what are the company plans for their future wireless communication systems in intralogistics and which are the related potentials and challenges identified?

To bring the focus on this issue employees of various companies in Germany were asked about the use of wireless communication technology in intralogistics, in order to carry out a qualitative survey. The aim of this study is in the first place to determine which communication technologies are already being used in the companies, what requirements exist, how these are defined by the responsible employees in the companies and what challenges are usually recognized or expected during the implementation of 5G or other new communication technologies.

In the following is an overview on the procedure of the conducted study as well as the results of the expert interviews and company survey. The paper concludes with a summary and an outlook.

2 Systematic Procedure

The study consists of two parts. The first part is based on interviews with six experts from science and business in the field of intralogistics and communication technology. These experts were interviewed for about an hour in autumn 2021. They provided assessments and evaluations on the internal use of wireless communication in general and 5G in particular. They were also asked to evaluate the current use of 5G in intralogistics. The results are presented in Sect. 3.

The information gained from the first part served as a base for the questionnaire of the survey in the second part, which took place between the end of 2021 and spring 2022. In this part, a total of eight employees of various German companies of different sectors (e.g. logistics provider, manufacturing industry) and sizes (from small and medium-sized companies to large concerns) were interviewed. The interviewees have insights into intralogistics systems of their companies and hold for example positions like Vice President Global Innovation, Project Manager Logistics, Managing Director Logistics or CIO Innovation and IT. They answered a questionnaire consisting of eight questions about the use of existing wireless communication systems in their intralogistics as well as the current and future needs and 5G. The question about the challenges with existing wireless communication technology in general aimed to identify the need for replacing the infrastructure with other technologies. In addition, the survey was conducted with two further professionals from science and business who have insights into several companies as consultants in order to validate the results of the company survey.

The answers of the questionnaire were partly generalized to anonymize the data in order to avoid identification of the interview partners. The evaluation of the surveys was based on a qualitative content analysis according to Mayring [10]. In this method, the answers are assigned to different categories. These categories are then grouped into superordinate clusters. The number of responses belonging to a category is then determined. The survey results are analyzed by comparing the categories within the clusters. The results are shown in Sect. 4.

3 Results – Expert Interviews

The main key aspects from the conversations with the experts were identified and summarized. The most important results are below:

1. 5G is not yet widespread in smaller companies

The interviewed experts agreed on the diagnosis that 5G so far has not found its way into German companies, especially among smaller companies it is often not even considered. In addition to the relatively high costs, it is often the case that due to limited application spectra in companies the new technology seems far from being fully exploited even though through more efficient data use and the capability of processing many more data is also not considered as urgently needed to maintain the market position.

2. For companies, the benefit of implementing 5G must be illustrated by areas of application

It is important to first identify the areas of application for this technology in order to be able to assess the benefits for companies. Here it seems particularly important to provide examples which take into account the benefits for companies in their current day to day work process as well as to generate an outlook on the future use and advantage of a system integration towards an interactive common communication system covering most of the processes in any particular company.

3. Uses Cases for 5G may be defined by the failure of conventional wireless communication systems

The establishment of a new communication standard is often only pursued in companies when there are indications that the existing systems are reaching their limits in the medium and long term and thus threatening to restrict the company's operability. It is therefore important for the industry to determine the limits of the existing systems in order to find out where and when, for example, WLAN is no longer sufficient to enable efficient communication. Herein use cases for 5G may even be defined by the failure of the conventional systems such as WLAN.

4. The decision on which communication technology to use is based on a cost-benefit analysis

The intralogistics industry fundamentally lacks use cases for 5G beyond AGV´s, since the conventional wireless communication systems are in place and sufficient to operate the standard processes and production. There is overall no urge to change to 5G unless the company is having the possibility to finance the change and are seeing it as an investment into the future competitiveness.

Taken together, the experts believe a general interest in using new technologies such as 5G. Large companies are planning to implement 5G in the future to enhance efficient utilization of capacities. Also there are pilot projects where 5G has already been implemented. However such a trend is not emerging among smaller and medium-sized companies. In addition, commercial solutions for the individual configuration of the networks, which are oriented towards the requirements of the companies or even tailor-suited technical solutions are only available to a limited extent.

4 Results – Company Survey

In this section, the results of the company survey, which were conducted in the second step, are presented. First the questions are given with the corresponding summarized answers of the company representatives. The results are then validated by two further surveys of consultants from science and industry.

1. For which applications do you use battery-powered/ wireless systems in intralogistics?

In the first question all interviewed persons stated that their companies use industrial trucks such as forklifts. Some of them already use AGVs (40%; 3/8). Many use scanners

(90%; 7/8) and terminals (60%; 5/8), for example to document the movement of goods. In addition, pick-by-vision equipment, inventory drones and printers are occasionally used.

2. What use cases do you have, in which data communication has a very high priority for safety in intralogistics?

This question was answered on the one hand from the aspect of safety measures, i.e. the machine is not allowed to be dangerous for people and the environment, and on the other hand from the aspect of reliability. Regarding safety, AGVs and general human-machine interactions are examples for the need of reliable and high-throughput data transfer. When it comes to reliability, companies are concerned with ensuring the functionality of processes, such as the availability of device connectivity in the area of a management system in outdoor areas.

3. Which wireless communication technology do you use in intralogistics?

All companies of the respondents use WLAN in intralogistics. In addition to WLAN, many use mobile communications (60%; 5/8) and some of them Bluetooth (30%; 2/8). Some applications use multiple technologies in combination. An example for this is scanners connected to WLAN indoors and 4G outdoors.

4. What challenges do you currently face regarding the existing communication infrastructure in intralogistics?

In the context of WLAN, challenges mentioned concerned mainly the areas of coverage, availability and capacity overload. 60% (5/8) addressed the general problem of wireless coverage, especially when existing warehouses are refurbished and extended. Furthermore, some see challenges in the wiring of WLAN access points and the associated infrastructure costs, the lack of flexibility and the general availability of networks. Challenges exist in terms of network accessibility around metals, liquids and outside warehouse or in outdoor applications. In some cases, an overload and failure of the network caused by a simultaneous enquiry of a high number of clients was reported as well as during peak-times when communication between interfaces excesses the capacities of the WLAN. In connection with wired solutions the low flexibility was identified as a main problem.

5. What are your requirements for the communication infrastructure in intralogistics?

For many companies of the survey participants, stable availability of the communication technology (80%; 6/8) is important. Also a network latency that is sufficient for the respective application (60%; 5/8). Other requirements were involving the transmission rate but also data security, number of devices, costs as well as use cases were important issues.

6. Where could your communication infrastructure in intralogistics reach its limits in the next three to five years?

In terms of future perspectives they stated that limits are in the expansion of production sites, the use of new technologies and the striving for a higher degree of automatization. The implementation of AGVs, inventory drones and automated warehouse technology were mentioned as examples.

7. Have you already dealt with the use of 5G?

60% (5/8) of the interviewed company representatives have examined the possible uses of 5G. Some of these companies (40%; 3/8) concluded that they are currently not planning to implement this technology, as they have not identified the need to do so and for financial reasons. 30% (2/8) reported they are using 5G in a test area or pilot project. Thus, no one is using 5G for daily operations as a wireless communication technology.

8. What potential future use cases of 5G do you have in mind?

The companies named numerous potential use cases that go beyond the use of AGVs. These include, for example, route optimization, localization tasks and processing of orders in real time, networking of machines and automation of warehouses. Furthermore, the use of new technologies such as inventory drones, video analytics as an alternative to scanners and augmented reality, for example in the field of order picking, was mentioned.

In addition to answering the questions in the questionnaire, general comments were also made on the use of 5G in industry. Amongst these the availability of hardware is important for the successful implementation of 5G. Currently there is a lack of equipment, such as access point routers based on 5G. Furthermore, it was criticized that the general infrastructure is insufficiently developed, for instance the companies are obligated to install 5G-masts in order to implement 5G.

While some use cases were described, there was a consensus that there is a lack of use cases that justify the introduction of 5G in intralogistics. This indicates that there is still the need for further advice on what the technology can be used for and services to enable implementation in a way that ensures uninterrupted operational functionality of the companies.

In addition to the eight company representatives, two further consultants were interviewed using the questionnaire for validation and to include a wider view on the topic. These two consultants also named industrial trucks, terminals and scanners as battery-operated systems and part of the wireless infrastructure usually found in intralogistics. Similarly to the results obtained beforehand the importance of reliability was emphasized, and stressed their role for proper operation. Like the company representants, the consultants mentioned some challenges when WLAN is applied as the main wireless communication technology since the coverage for the entire company premises and the associated infrastructure is difficult to achieve. Expensive installation and the lack of flexibility were considered problematic. In addition, the time for switching between access points and their limited capacities regarding the number of devices controlled

simultaneously are named as a weakness of WLAN. Other problems are the susceptibility to interference due to signal deflection, for instance steel processing. The outdoor use is also limited. These findings also coincide with the companies statements.

According to the consultants assessment of so far described use cases there is no such things as a "killer application" for 5G, but it is expected there will be positive effects because of the further development of the technology. However, for an appropriate use of 5G the availability of compatible devices is essential, but this hardware is currently not available. According to the consultants, there are no small modules that can be integrated in existing systems. Another major problem is that 5G-solutions currently are characterized by a very high energy consumption. In summary, it can be said that even according to the consultants intralogistics is not yet ready to use 5G on a daily base. The situation is different in research, where initial applications are being tried out in various projects.

To summarize, all interviewees for intralogistics in the industry are found to be aware of the new mobile communications standard 5G. They recognise that in the near future 5G or other new standards will need to be introduced to keep up with digitalisation and automatization in the German economy as a whole. However, most of the companies have not yet implemented or even considered 5G for their businesses, in some cases its potential use was evaluated but have refrained from implementing the technology in their company due to cost and adequate capacities provided by common wireless communication techniques such as WLAN. Many companies are currently in a waiting position, hoping for the further development of devices and commercially available systems solutions ready for use and tailor-fitted for their company in the next few years. Furthermore, a barrier for the implementation of 5G, is the lack of studies and reports quantifying added values for a realistic cost-benefit analysis. In general, the expected relevance of 5G in intralogistics was considered high in the next two to three years. Especially if the shortcomings of WLAN cannot be overcome, it should be examined whether the use of 5G is a better alternative. Based on the results of the studies conducted, the following principles have been derived to answer the research questions formulated at the beginning, which targeted the future use of 5G and the state of wireless communication technology applied in intralogistics:

1. WLAN is currently the technology of choice for wireless communication in intralogistics
2. 5G is still at an early stage in intralogistics and is not very widespread; there is a lack of soft- and hardware as well as non-existing or unfavourable cost-benefit analysis of use cases defined so far
3. When WLAN reaches its limits, 5G is seen as an alternative for the future
4. The greatest challenges of WLAN in intralogistics are the coverage of the locations and the access in outdoor areas; here 5G appears to be a suitable alternative

The findings of this study on the status quo of 5G in intralogistics coincide with the results of the studies by Franken et al. [3] and bitkom [2]. The former also mentioned that the added value is hard to quantify and soft- and hardware are not readily available. In bitkom [2] the use of WLAN and an insufficient expected added value were reasons against the use of 5G for interviewed industrial companies.

5 Conclusion and Outlook

In the field of intralogistics, a wide range of different wireless systems are used in companies. Devices such as industrial trucks, terminals and scanners are commonly used and dependent on wireless communication. Readily available communication technology is particularly important to companies in order to enable productive operations without problems. In general, the companies report that the wireless communication technologies currently in use are satisfactory, for intralogistics predominantly WLAN. However, the use of WLAN is challenging, especially regarding a stable and sufficient coverage over the entire surface of a production side and the outdoor area. In addition to this there are considerable costs associated to an extended WLAN due to the need to install a high number of access points. This complex communication systems result in restrictions regarding flexibility when production and logistics operation processes change. 5G offers the opportunity to overcome some of the challenges associated to the commonly used wireless communication systems.

The study presented in this paper gives a first insight into the situation and the attitude of German companies towards the use of 5G in intralogistics. It provides a basis for the development of a further quantitative study. The study shows that the use of 5G in intralogistics is still in its infancy, Problematic is an insufficient knowledge about new technologies and their potential. While transparent and easily accessible information as well as contact points in national institutions may be able to serve as sources of information to promote the new technology standard by revision of its possible use in companies as well as by presenting concrete use cases.

A forecast for the next three to five years may be deduced from the available data and stated options: Due to the increasing degree of automatization, it can be assumed that some companies will increasingly deal with the use of 5G in their intralogistics. However, widespread use is not expected. The speed of development also depends on the provision of 5G-capable software and use cases. In Germany, there are currently many ongoing publicly funded projects in which industrial applications are being piloted with the 5G mobile communications standard. The funding includes test fields that are accessible to companies to test their own applications, but also projects to pilot specific applications [11].

In addition, there are several institutions that have made it their goal to describe 5G applications in intralogistics, among other areas [12].

Acknowledgment. This work is part of a funded project “5G-trAAffic”, funding guideline “5G-Umsetzungsförderung im Rahmen des 5G-Innovationsprogramms” of the Federal Ministry for Digital and Transport (BMDV) with the funding code FKZ 165GK044A. This publication was also partly created within a project funding “AuLoKomp” of the Centre for Digitisation and Technology Research of the German Armed Forces.

References

1. Grotepass, J., Eichinger, J., Voigtländer, F.: Mit 5G zu neuen Potentialen in Produktion und Logistik. In: ten Hompel, M., Bauernhansl, T., Vogel-Heuser, B. (eds.) Handbuch Industrie 4.0, pp. 251–284. Springer, Heidelberg (2020). https://doi.org/10.1007/978-3-662-58530-6_106

2. bitkom: 5G in der Industrie (2019). https://www.bitkom.org/sites/default/files/2019-05/190515_bitkom_charts_pk_5g.pdf. Accessed 17 Mar 2022
3. Franken, M., Sörries, B., Stronzik, M.: Entwicklung von 5G-Campusnetzen in Deutschland. Netzwirtschaften und Recht **18**, 257–320 (2021)
4. Winkler, H., Zinsmeister, L.: Trends in digitalization of intralogistics and the critical success factors of its implementation. BJO&PM **16**, 537–549 (2019). https://doi.org/10.14488/BJOPM.2019.v16.n3.a15
5. Fottner, J., et al.: Autonomous systems in intralogistics-state of the art and future research challenges. Logist. Res. **14**, 2 (2021)
6. Boshoff, M., Miro, M., Kuhlenkötter, B.: Anforderungen für den Einsatz von 5G in Produktionsumgebungen. Zeitschrift für wirtschaftlichen Fabrikbetrieb **117**, 84–87 (2022). https://doi.org/10.1515/zwf-2022-1010
7. Evcenko, D., Kett, H.: 5G-Nutzenbewertung. Zeitschrift für wirtschaftlichen Fabrikbetrieb **117**, 88–93 (2022). https://doi.org/10.1515/zwf-2022-1014
8. Bundesnetzagentur: Übersicht der Zuteilungsinhaber: für Frequenzzuteilungen für lokale Frequenznutzungen im Frequenzbereich 3.700–3.800 MHz (2022). https://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Sachgebiete/Telekommunikation/Unternehmen_Institutionen/Frequenzen/OffentlicheNetze/LokaleNetze/Zuteilungsinhaber3,7GHz.pdf?__blob=publicationFile&v=9. Accessed 17 Mar 2022
9. Bundesnetzagentur: Antragsverfahren für lokale Frequenznutzungen im Bereich 3.700–3.800 MHz (2020). https://www.dihk.de/resource/blob/25190/6191b0e5596e792151b700611e948400/praesentation-gawron-2020-06-23-data.pdf. Accessed 17 Mar 2022
10. Mayring, P.: Qualitative Inhaltsanalyse, vol. 14, pp. 159–175. UVK Univ.-Verl. Konstanz (1994)
11. Bundesministerium für Digitales und Verkehr. 5G-Mobilfunk. 2022. https://www.bmvi.de/DE/Themen/Digitales/Frequenzen-Mobilfunk-und-Digitalradio/5G/5g.html. Accessed 1 Jul 2022
12. VDMA: Wireless Communications for Machines. https://www.vdma.org/wireless-communications-for-machines. Accessed 1 Jul 2022

Concept for a Revolution of Public Transport

Wolfgang Echelmeyer, Tuan Nguyen(✉), and Mert Mete(✉)

Reutlingen University, Alteburgstraße 150, 72762 Reutlingen, Germany
wolfgang.echelmeyer@reutlingen-university.de

Abstract. Public transport causes in rural areas high costs per passenger and kilometer as the frequency of scheduled busses is low and therefore, many people avoid using public transport. With the trend of moving from urban regions to countryside individual traffic will further increase. To tackle issues of emissions, mobility for young and elderly people and provide economically meaningful public transport a new concept was elaborated in Germany. This consists of (partly) autonomous shuttle busses which are remote controlled. For implementation rural districts of Germany have worked together and set up a three-phase plan consisting of a project with public funding, a highly frequent used pilot region and industrial partners with the commitment and possibilities for necessary investments. The concept promises economical value with respect to installation, service and maintaining costs, it leads to lower barriers for public transport of young and elderly people and ultimately reduces emissions and congestions.

Keywords: Public transport · Business model · Autonomous transport · 5G

1 Introduction

According to the most recent statistics on public transport in Germany nearly 11,7 billion passengers were using either bus or train in 2019. With the exception of the last two pandemic years the demand for public transport never stopped increasing over the last 10 years [1]. To handle this demand, in Germany 96,000 people alone are working in the sector of public transportation (specifically as drivers). It is prognosed that due to the demographic situation in Germany 40,000 of this 96,000 people will be retired leaving a hardly fillable gap. At the same time the goal of public transport companies until 2030 is to increase the number of busses and trains on the road and rail up to 1/3 to make public transport more attractive [2].

A study from 2019 about urban vs. rural public transportation shows that citizens in rural areas are 2.8 times less likely to use public transportation than urban citizens. This is due to the lack of public transport infrastructure commonly found in rural areas. One reason for this deficit are the high costs associated with maintaining such inefficient transport routes. Urban areas have a higher density, thus more people can be transported on a shorter distance while in rural areas the passengers per kilometer are significantly less resulting in higher costs [3].

To tackle these aforementioned problems this paper introduces a concept and vision which solves the inefficient cost ratio of public transport in urban and especially rural

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 489–494, 2023.
https://doi.org/10.1007/978-3-031-16281-7_46

areas. In this paper we first introduce the vision and its planned phases of implementation. Then we present the identified challenges that need to be overcome before we report about the solution and the purpose of the vision. Finally, we discuss the outlook and the next steps to accomplish the vision.

2 Vision

The here presented concept for a new local public transport follows a vision of a far more economical and user-friendly system than it is today, especially in rural areas. In this vision, autonomous shuttle buses operate at frequent intervals in the countryside and serve as feeder services to both local centers and main transport axes. Additionally, remote control of vehicles by using 5G can achieve a higher degree of automation so that the vision can be implemented economically even in less populated areas.

The concept, which was elaborated in 2021, is more than a vision. It contains a business model that makes it possible to make statements about the economic efficiency depending on input key figures and parameters such as installation costs and desired degree of automation. It shows a vision where we evolve from non-automated shuttle buses in public transport today to shuttle buses with a certain degree of automation. The core element of the overall vision is the aforementioned (5G) radio remote control of (otherwise autonomously driving) vehicles in critical situations. This means that each individual vehicle no longer needs to have an individual driver on board. This is because 5G provides a technology with such low latency and such high data transmission rates that several vehicles can be monitored simultaneously by one person from a remote-control center and, if necessary, safely intervene remotely (see Fig. 1). The concept also allows to develop new approaches for public transport e.g., on-demand transfers or video chat with remote controllers to inform about delays.

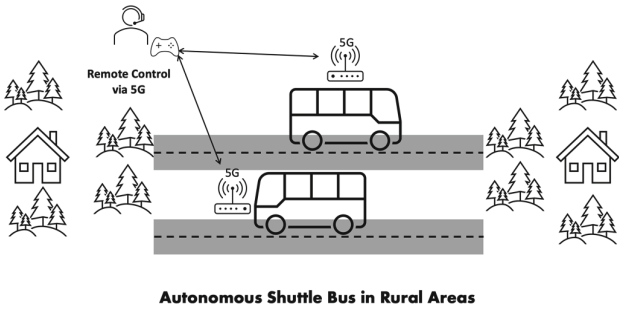


Fig. 1. Concept of the vision in rural areas

The concept can develop into a highly efficient system that requires only one-third to one-fifth of the current number of drivers and thus eliminates a significant part of the kilometer-dependent costs. A cost comparison conducted by Bösch [4] (Fig. 2) shows how much the costs per passenger kilometer of different transport modes and vehicle types in rural and urban areas would be reduced if the labor costs to drive the vehicle

were completely eliminated. According to this study, the operating costs of a bus could be reduced by more than 50% from 0.53 Swiss Franc (CHF) to 0.24 CHF in urban areas if you compare autonomous (Auto) vs. conventional (Conv) modes of transport.

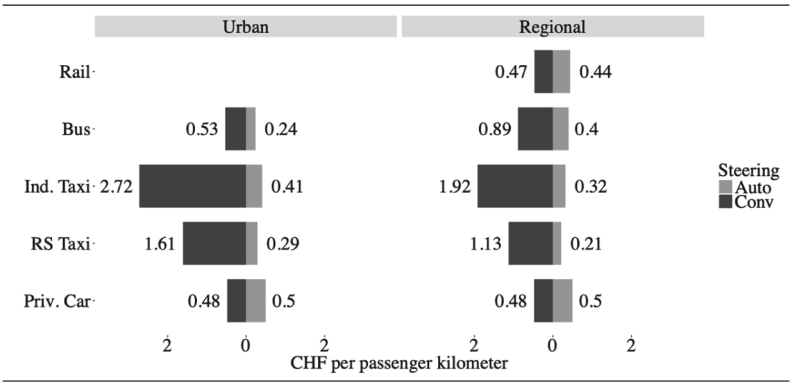


Fig. 2. Cost comparison of different modes with (*Auto*) and without (*Conv*) autonomous vehicle technology according to Bösch [4]

The concept was elaborated by three German districts (Reutlingen, Sigmaringen and Zollernalb) as well as the city of Reutlingen, who are applying for implementation funding in the 5G innovation competition of the Federal Ministry of Transport as a consortium in cooperation with industrial companies. The concept is the basis of a plan that is to be implemented in three phases, where some of them will be worked on in parallel:

1. Industrial research with the aim of enabling radio remote control of buses via a control centre. The buses should mainly drive autonomously and the staff in the control centre should only intervene in critical situations. One person in the control centre will monitor several vehicles and control them individually if necessary. The new RTunlimited industrial and commercial park in Reutlingen was chosen as the environment for technical tests. 5G infrastructure can be used there, which is currently being setup independently of the concept and its implementation.
2. In one pilot location autonomous bus shuttles and 5G infrastructure with technology available on the market will be used for testing under real environmental conditions. The Outletcity Metzingen was selected for this purpose, because economically viable operation can already be expected there due to the circumstances and the expected number of users. This pilot location serves to validate operating and organisational concepts and, furthermore, as an experimental field for further increasing the degree of automation of the vehicles.
3. Planning and calculation of further pilot sites in rural areas, depending on the key outcome of the technology and business models and within the pilot region in Metzingen (installation costs, degree of automation, operating costs). For this purpose, the developed calculation model is used and continuously developed. This phase

ensures that the results of the project will be made available for other rural areas beyond the funding period.

3 Challenges for Future Vision of Public Transport

In order to be able to implement the above-mentioned vision of public transport, three major challenges must be overcome. Technical preconditions must be created, economic and legal hurdles must be removed.

- Technically preconditions: Autonomous minibuses are not a hurdle as they are already available on the market and in use at some pilot locations in Germany. The acquisition costs are currently still relatively high, but significantly lower purchase prices can be expected in the coming years due to series production of the vehicles. The necessary 5G infrastructure and its use are already being planned at the first test and pilot locations in Reutlingen and Metzingen.
- Economic hurdles: Today, from an economic point of view, it is noticeable that up to two thirds of the costs of existing call-bus systems are pure provision costs. This is precisely why this is an ideal field of application for 5G technology. It considerably reduces personnel requirements and personnel costs on site by making it possible or easier to provide new services.
- Legal hurdles: This seems to be the biggest of the aforementioned challenges. It lies in the fact that the law currently does not allow fully autonomous operation without a (reserve) driver in the vehicle. Nor does it seem foreseeable that this will be possible at normal traffic speeds in the near future. The concept therefore relies on the idea of remote control, which combines (partially) autonomous driving with continuous monitoring and the possibility of intervention from a central control centre at any time. The researchers at Reutlingen University, who developed the concept, believe that after discussions with industry and politicians, there is a good chance that not only the technical but also the legal hurdles can be overcome with this combination - at least for this project. Then it could be shown that the desired (personnel) cost reduction can be achieved, which in principle also makes a permanent implementation economically possible.

4 Solution and Purpose

The technical task of the project is to test the monitoring and remote control of an autonomous minibus using the 5G network. With its high transmission rates of 20 GB per second (at least 40 times more than 4G) and latency times of 1 ms (100 times faster than 4G), the 5G network provides the technical basis for this. This is because the operator can capture the vehicles' surroundings and the traffic situation from the control centre using high-resolution camera images with virtually no delay and control the vehicle. In future, remote control in real time should also enable higher speeds ($>$30 km/h) without greater risk.

The concept displays the connection between improved economic efficiency of public transport, improved mobility, decrease of congestion and decrease of energy consumption. Improved economic efficiency was already addressed in one of the sections above by replacing many drivers by one person with remote control. But furthermore, the economic benefits are particularly great in rural areas. Roads with lower traffic volumes are easier to control as there is less infrastructure e.g., traffic lights, circular traffic such that it is possible to implement trouble-free shuttle operation quicker and with less effort. Hence, it is possible to offer public transport with a higher frequency for the same costs. This leads to improved mobility especially for young people with no driving license or elderly people, making mobility possible for otherwise non-mobile persons. This will even be of higher impact as according to the Federal Institute for Research on Building, Urban Affairs and Spatial Development (Bundesinstitut für Bau-, Stadt und Raumforschung, BBSR), there has been a renewed trend towards living in the countryside since 2014, as property prices in the city have risen steadily and are no longer affordable for many people. Therefore, without new concepts the individual traffic, already accounting for 70% of all journeys, will increase further [5]. The concept will increase use of public transport, reduces private transport, thus decrease environmental impact of congestion and the associated increased energy consumption of vehicles being in congestions. The technically optimized and automated driving processes is a further impact on reducing energy consumption. The shuttles planned for the concept are electrically powered resulting in no local emissions. According to experts, the last aspect for reduction the of energy consumption is, if all vehicles are fully automated, traffic flow is smoother and will reduce emissions [6].

Last but not least the concept also shows that 5G technology creates added social value for the population in rural areas.

5 Outlook

The pilot region under consideration is particularly suitable for testing the autonomous shuttle buses. Against the background of the planned Neckar-Alb regional light rail system, the additional services represent an efficient supplement and combination for rural public transport. The regional light rail system includes various extension and new construction routes in the Neckar-Alb region with the districts of Reutlingen, Tübingen and Zollernalb with around 900,000 inhabitants. Autonomous shuttles can therefore also be considered as feeders to the regional light rail system in the future. The defined pilot locations are located in the direct or extended catchment area of the regional light rail concept. The combination of shuttle buses and regional light rail is suitable to increase the benefits of both systems. The concept also shows possible solutions for other locations where the infrastructure is not yet sufficient. It´s needless to say, that 5G is not fail-proof so if the concept is being deployed into a real environment many different factors and problems need to be addressed and being responded accordingly (e.g. what happens if the vehicle is inside a tunnel). Due to the concept only being at the stage of a vision these hurdles will be tackled in a later stage of implementation.

References

1. Statistisches Bundesamt: Anzahl der Fahrgäste im Linienverkehr mit Bussen und Bahnen in Deutschland von 2010 bis 2021. https://de.statista.com/statistik/daten/studie/1233244/umfrage/fahrgaeste-im-linienverkehr-mit-bussen-und-bahnen-in-deutschland/. Accessed 03 June 2022
2. Verband Deutscher Verkehrsunternehmen: Personal- und Fachkräftebedarf im ÖPNV. https://www.vdv.de/personal-und-fachkraeftebedarf-im-oepnv.aspx. Accessed 03 June 2022
3. Tao, X., Fu, Z., Comber, A.: An analysis of modes of commuting in urban and rural areas. Springer Verlag (2019)
4. Bösch, P.M.: Autonomous Vehicles - The Next Revolution in Mobility. ETH Zürich, Zürich (2018). https://doi.org/10.3929/ethz-b-000296870
5. Bundesinstitut für Bau-, Stadt und Raumforschung: Landflucht? Gesellschaft in Bewegung, BBSR Verlag, Berlin (2016)
6. Greenblatt, J.B., Shaheen, S.: Automated vehicles, on-demand mobility, and environmental impacts. Curr. Sustain. Renew. Energy Rep. **2**, 74–81 (2015). https://doi.org/10.1007/s40518-015-0038-5

Human-in-the-Loop Control of Haptic Devices: Now and the Future

An Intelligent System for Human Intent and Environment Detection Through Tactile Data

Gianluca Laudante(✉) and Salvatore Pirozzi

Università degli Studi della Campania "Luigi Vanvitelli", Aversa, Caserta, Italy
gianluca.laudante@unicampania.it

Abstract. The tackled application aims at demonstrating the effectiveness of using tactile sensors for multiple purposes at the same time. In particular, tactile data are exploited for: the estimation of a wire diameter by using a previously trained classifier; teaching the robot new wire routing trajectories in case of unknown grasped wires; stopping the autonomous execution of previously learned trajectories to avoid damages due to possible wire entanglements.

Keywords: Tactile sensors · Wire manipulation · Intelligent system

1 Introduction

Tactile sensors represent a key enabling technology to improve the robot capabilities in the execution of complex tasks. For example, grasping and manipulation tasks can be successfully executed only by knowing specific geometrical and physical features of the objects, together with an estimation of interaction forces/torques between the robotic system and the environment. In recent applications, the robotic systems are also often used to implement safe and efficient physical Human Robot Interaction tasks, which requires the knowledge of interaction forces and contact locations in order to perform cooperation and comanipulation tasks and to limit damage from accidental impacts. To this aim the robots are more and more frequently equipped with sensorized grippers, capable to reproduce an artificial sense of touch, trying to imitate humans, which can perform complex tasks thanks to their sophisticated tactile perception [1,2].

Recently, many researchers started to present results about touch sensing solutions applied to daily life applications, based on both model based and machine learning approaches. In [3], authors present an application to distinguish materials on the basis of friction, by using a Naive Bayes classifier. Bandyopadhyaya et al. [4] uses a piezoresistive tactile sensor to distinguish objects of different softness using Decision Tree and Naive Bayes methods. In [5], authors presented a solution for the discrimination of object curvature using a sparse tactile sensor array. Da

This work was partially supported by the European Commission within the H2020 REMODEL Project (no. 870133).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 497–506, 2023.
https://doi.org/10.1007/978-3-031-16281-7_47

Fonseca et al. [6] proposed a combination of machine learning methods and fuzzy logic controllers to improve the grasp stability through tactile data for an under-actuated hand. An interesting feature of tactile sensors is that they can be used not only for recognition and classification strategies as discussed above, but also for control purposes, such as grasping and manipulation [7–10]. The interested reader can deepen the topic of tactile sensing for dexterous manipulation by referring to the review papers [11,12], and references therein. As detailed in the following, in this paper the tactile sensor will be exploited also as teaching system, as alternative approach to commonly used solutions [13,14].

The main objective of this paper is to demonstrate that the tactile sensor, developed by the authors and detailed in [15], can be used to tackle complex tasks by combining machine learning and model based methods applied to the same tactile data. In previous papers, the authors demonstrated the effectiveness of the proposed technology for the shape recognition of grasped wires during manipulation tasks devoted to the switchgear assembly [16] and a solution for the classification of wires on the basis of their diameters [17]. In this work, past and new methods are combined in a more articulated solution, where teaching phases and task execution phases are automatically managed on the basis of object classification. During the teaching phases an human operator interacts with the robotic system by exploiting the tactile data as guidance system, while during the execution phases the tactile data are used to avoid damages in case of unintended environmental interactions. A classification method trained on the basis of tactile data supervises if the grasped object needs a teaching phase or can be directly used for an execution task. In the next sections the tackled task is detailed together with the proposed approach. Finally experiments are reported to demonstrate the effectiveness of the implemented solution.

2 Interaction Task

As discussed, the objective is to demonstrate how it is possible to exploit the same sensor to execute a complex task. To this aim, a single tactile sensor is used to accomplish different phases in an example application: routing a wire following a certain path depending on the diameter of the wire.

The whole task is resumed in the flowchart in Fig. 1(a). Initially, the robot moves to a fixed starting position to grasp a wire. Then, an estimation of the wire diameter is obtained and it is used to search for a trajectory in a database. In case there is no saved trajectory, an operator, by pulling the wire, teaches the robot the path to follow the next time a wire with the same diameter is grasped. Instead, if a trajectory already exists, the robot executes the wire routing as it has been taught and stops the motion in case a force on the wire is detected, which could be due, for example, to an entanglement during the trajectory execution.

During the task, the tactile sensor is used for three different purposes:

1. Estimate the diameter of the wire;
2. Compute the command for moving the robot during the teaching phase;
3. Stop the robot if the wire gets entangled during the autonomous routing.

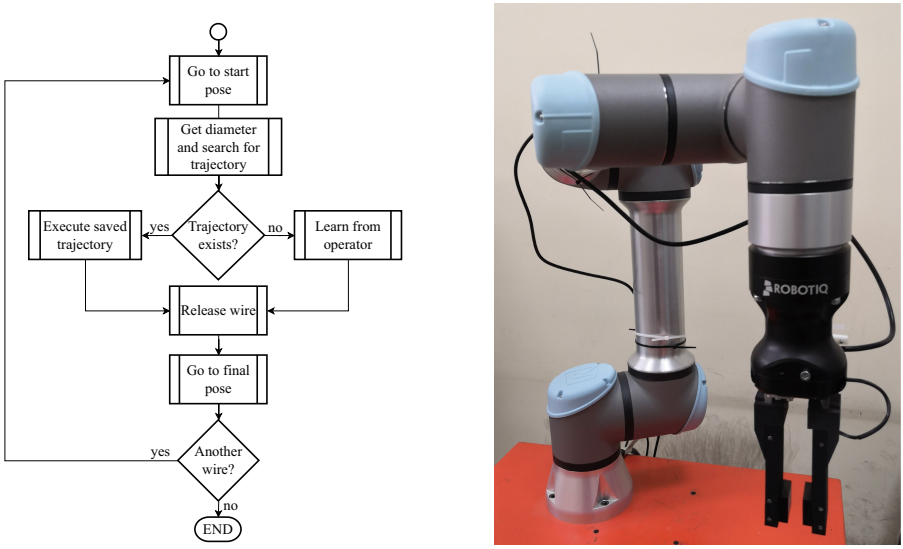


Fig. 1. Task flowchart (**a**) and hardware setup: robot, gripper and fingers (**b**).

2.1 Task Robotic System

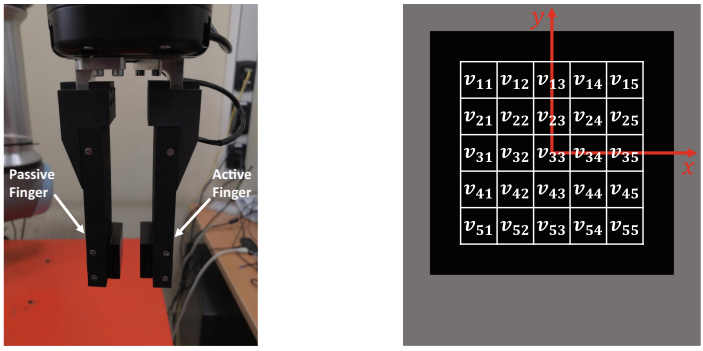
The robotic system consists of an Universal Robot UR5e manipulator equipped with a Robotiq Hand-E gripper that has a couple of custom designed fingers: an active finger equipped with a tactile sensor and a passive one (see Fig. 1(b)).

Tactile Sensor. The active finger integrates the tactile sensor detailed in [15]: it consists of 25 photo-reflectors integrated into a standard Printed Circuit Boards (PCB) and suitably assembled with a deformable pad made of silicone. The optoelectronic devices are organized in a 5×5 matrix, with a spatial resolution equal to 3.55 mm and a sensitive area of about $21 \times 21\,\text{mm}^2$. The photo-reflectors are constituted by infrared LEDs and photo-transistors optically matched. The LEDs are driven by adjustable current sources and the photo-transistors signals are acquired by a microcontroller equipped with 25 low-noise A/D channels with a 12-bit resolution. The silicone pad transducers the external contacts into deformations measured in a discrete number of points by the optical devices. The measurements correspond to the voltages available on the photo-transistors on the basis of light emitted by LEDs and reflected by the bottom side of the deformable pad. Table 1 reports the main features of the sensor, but the interested reader can find additional details in [15]. The passive finger is equipped with the same deformable pad as the active one, but without the optoelectronic components. Figure 2(a) reports a detail of the fingers mounted on the gripper.

A suitable ROS (Robot Operating System) node has been developed to make the tactile data available from the microcontroller. Initially, the node acquires

Table 1. Main features of the active finger.

Taxels number	Sensing area	Spatial resolution	Sampling frequency	Response time
25 (5×5)	$21 \times 21\,\text{mm}^2$	3.55 mm	500 Hz	<0.01 s



(a) (b)

Fig. 2. Active/Passive fingers couple (**a**) and taxel matrix of the tactile sensor (**b**).

the voltage offsets from all photo-transistors when the sensor is in rest condition. These offsets are the mean value of the first 50 received voltage samples. Then, in normal working conditions, the node makes available, at a frequency of 500 Hz, the tactile map corresponding to acquired voltages cleared from initial offsets, denoted as v_{rc} in the following. Also details about the software can be found in [15]. Figure 2(b) shows the taxel distribution in the 5×5 matrix and the sensor frame positioned in the center of the sensing area, by highlighting for each taxel the corresponding voltage signal v_{rc}, defined by exploiting its row and column indices within the matrix.

2.2 Task Program Architecture

The complete program architecture for the proposed task, represented with a block scheme in Fig. 3, is developed using ROS,[1] an open-source software framework for robotic applications. Each block in the scheme is detailed in the following.

Diameter Classifier. An estimation of the diameter of the grasped wire is needed for the task execution. This is achieved by a suitably trained classifier based on a machine learning algorithm, detailed in a previous paper [17] by the authors. To resume, the diameter is estimated by a classifier, consisting in a feedforward neural network, able to recognize 6 different diameters (i.e., [1.5,

[1] https://www.ros.org/.

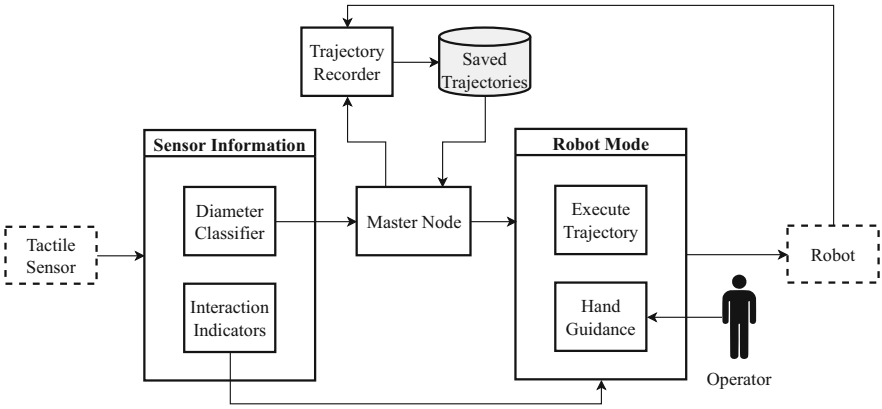


Fig. 3. Block scheme of the program architecture.

2.0, 2.5, 3.0, 3.5, 4] mm). The inputs to the neural network are 26 values, i.e., the 25 voltage signals v_{rc} of the tactile sensor and the relative distance between the gripper fingers. The node implements a ROS service[2] that returns the estimated diameter of the currently grasped wire.

Interaction Indicators. The 25 tactile signals are also used for the computation of three indicators, required by other components of the program. These indicators are all based on a comparison between the current tactile map, i.e., at time t, and the tactile map right after the initial grasp, i.e., at time t_0.

The first indicator is the sum of the absolute values of the variations in the tactile map:

$$i_{stop} = \sum_{c=1}^{5}\sum_{r=1}^{5}(|v_{rc}(t)| - |v_{rc}(t_0)|) \tag{1}$$

This is a generic indicator to detect an interaction with the environment.

The other two indicators are used to move the robot according to the force exerted by the operator on the grasped wire during the teach phase. In particular, one is used for the translation velocity of the end effector and the other for the rotational one, since the first is related to a shear force and the second to a rotational torque. Starting from the first indicator, it is computed as:

$$i_{transl} = \sum_{c=1}^{3}(\sum_{r=1}^{5}(v_{rc}(t) - v_{rc}(t_0))) + \sum_{c=3}^{5}(\sum_{r=1}^{5}(v_{rc}(t) - v_{rc}(t_0))) \tag{2}$$

that is the sum of the total variation in the left part of the sensor, i.e., columns 'c' from 1 to 3, and the total variation in the right part, i.e., columns from 3 to

[2] http://wiki.ros.org/Services.

5. The choice for this indicator comes from how the tactile signals respond to shear forces applied in the x direction of the sensor frame (see Fig. 2(b)), due to the asymmetry of the optoelectronic components. In fact, when a force is applied in the positive x direction the tactile signals increase, while they decrease when the force is applied in the opposite direction. Since the sign of all the variations is the same, the total variations of the left and right parts are summed.
The second indicator is similar to the first one and it is computed as:

$$i_{rot} = \sum_{c=1}^{3}(\sum_{r=1}^{5}(v_{rc}(t) - v_{rc}(t_0))) - \sum_{c=3}^{5}(\sum_{r=1}^{5}(v_{rc}(t) - v_{rc}(t_0))) \tag{3}$$

In this case the total variations of the left and right parts of the sensor are subtracted rather then summed. The choice of this indicator directly derives from the observation of how an operator would act on the grasped wire so to rotate the robot end effector. The most intuitive action is to pull the portion of the grasped wire protruding out one side of the sensor while pushing the one on the other side. As a consequence, this leads to opposite variations in the two sides of tactile sensor, i.e., positive in the "pushed" side and negative in the "pulled" one. Since in this case the variations have different signs, the total variations of the left and right parts are subtracted.

Execute Trajectory. This component deals with the autonomous routing of a wire on the basis of a previously taught trajectory and for point to point motion of the robot end effector. It is a ROS action server[3] that accepts a trajectory as a goal and returns a boolean value and a trajectory as result. The latter is important when the routing execution is halted after the detection of an external force applied on the wire, i.e., the i_{stop} indicator exceeding a fixed threshold. In this case, in fact, the node indicates that the task has not been completed and returns the portion of the trajectory not yet executed. In this way, an operator can choose to continue the task from the current position or to abort it.

Hand Guidance. This component transforms the interactions of the operator with the grasped wire in robot movements (in the 2D plane). This node, in fact, converts the indicators in Eqs. (2) and (3) in velocity commands for the robot. In particular, the conversion consists in applying a saturated dead-zone function, as the one in Fig. 4, to the indicators. It is useful to regularize the robot behaviour during the teaching when the indicators oscillate around zero, to limit the robot maximum velocity and to increase/decrease the sensibility of the robot velocity to the operator actions by changing the slope in the dead-zone function.

The node is implemented as a ROS action server accepting, as a goal, a target position and a timeout value. The returned result is a boolean value equal to *true* if the target position has been reached before the timeout, and $false$ otherwise.

Figure 5 shows two sequences representing the operator guiding the robot in both translations (Fig. 5(a)) and rotations (Fig. 5(b)) movements.

[3] http://wiki.ros.org/actionlib.

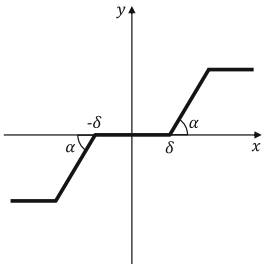
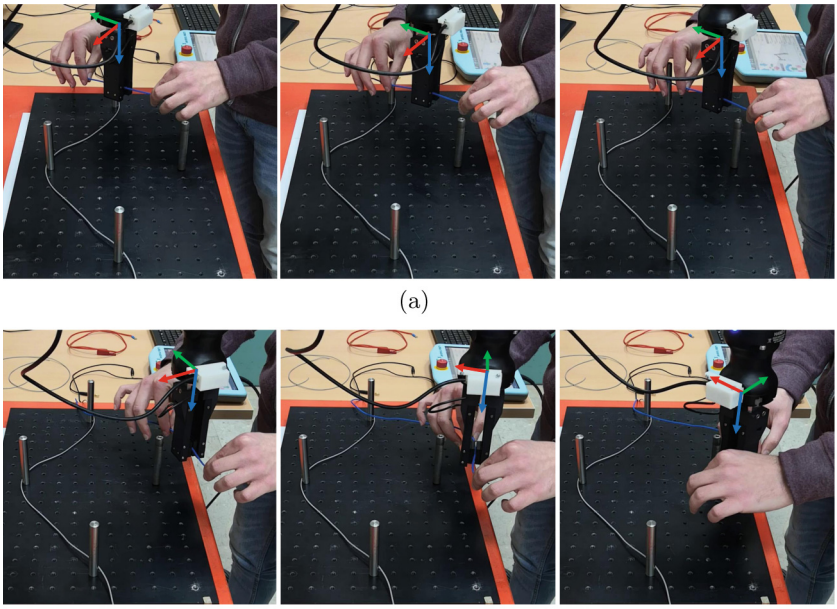


Fig. 4. Saturated dead-zone function.



(a)

(b)

Fig. 5. Operator guiding the robot in a translation (**a**) and in a rotation (**b**).

Trajectory Recorder. While the operator is guiding the robot, the executed trajectory needs to be "learnt" by the robot in order to be then reproduced. This is in charge of the *Trajectory Recorder* node, that is implemented as a ROS node offering two services: one to start the recording and one to stop it. The latter accepts a boolean flag to choose whether the recorded trajectory must be saved or not.

Master Node. This represents the "task executor", i.e., the component that controls all the operations. It starts the tasks by sending a two-points trajectory to the *Execute Trajectory* action server, in order to move the robot in a fixed start

pose. Once the end effector is in the desired pose, the gripper fingers are periodically closed until a contact is detected and the node asks the *Diameter Classifier* for the estimated diameter. Hence, knowing the diameter, the existence of the corresponding trajectory is checked. In case the trajectory does not exist, the node starts the recording by calling the corresponding service of the *Trajectory Recorder* node and sends a goal to the *Hand Guidance* action server. Once the teaching operation has been completed (or the timeout has been reached), the *Master Node* stops the recording asking the responsible component to save the trajectory (or discard it in case of timeout). Instead, if the trajectory already exists, it is loaded from the database, scaled in velocity if requested, and sent to the *Execute Trajectory* action server. In both cases, once the trajectory has been completed, the node opens the gripper fingers and commands the robot to reach a fixed final pose. Finally, the operator is asked if there is another wire to be routed and, in case, the procedure re-starts.

3 Experimental Results

This section reports an example test for showing the functionality of all the components in the program detailed above. The task consists in routing five electrical wires, with two having a diameter of 2.0 mm, two with a diameter of 2.5 mm and the last one with a diameter of 4.0 mm. At the beginning of the task no trajectory is in the database, so the operator teaches the robot the desired trajectories. The teaching phases are reported in Fig. 6, where the translational and rotational velocity commands derived from the tactile indicators are reported in blue, and the robot end effector pose in terms of x^b coordinate and angle θ_z^b about z axis w.r.t. the base of the robot are reported in red. After the teaching for each wire, the robot executes autonomously the routing for the remaining two wires following the correct trajectories. The pose of the robot during the autonomous routing is reported in the top graphs in Fig. 7, where the red line is the x^b coordinate of the end effector and the dashed blue line is the angle θ_z^b. In these graphs there are time intervals (i.e., [149, 154] s, [183, 187] s, [192, 196] s) during which the robot keeps the position. These intervals start in correspondence of peaks in the indicator i_{stop} reported in the bottom graphs of Fig. 7 and result from the operator pulling the wire during the trajectory execution, simulating an entanglement. The threshold for this test has been set to 0.12 V in order to be at least one order of magnitude above the sensor noise and to avoid that the weight of the wire itself is detected as a contact, stopping the execution. This value is tunable depending on the application and it can be reduced (taking into account the sensor noise) or increased to make the robot more or less sensible, respectively.

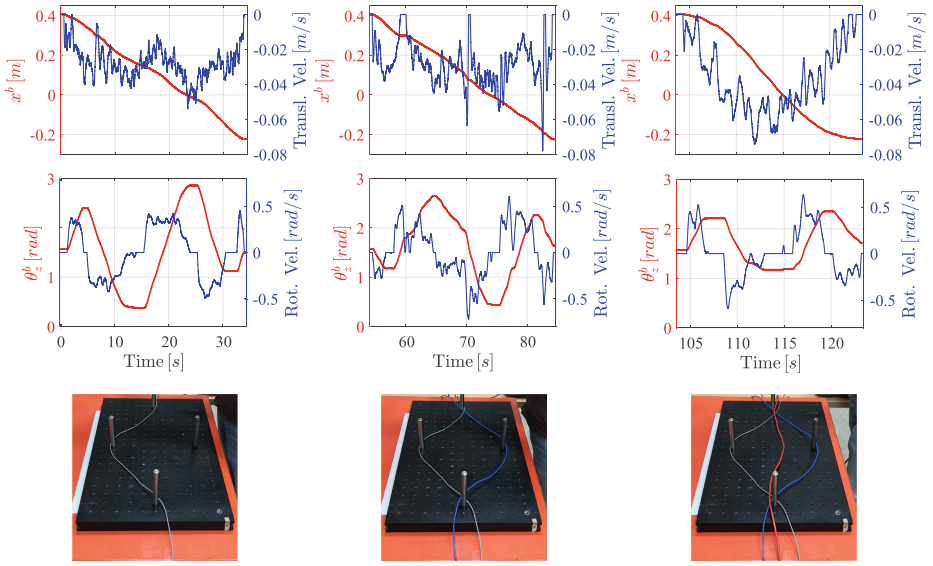


Fig. 6. Commanded robot velocities and end effector pose (x^b and θ_z^b) during the teach phase for wires with diameter of 2.0, 2.5 and 4.0 mm from left to right, respectively.

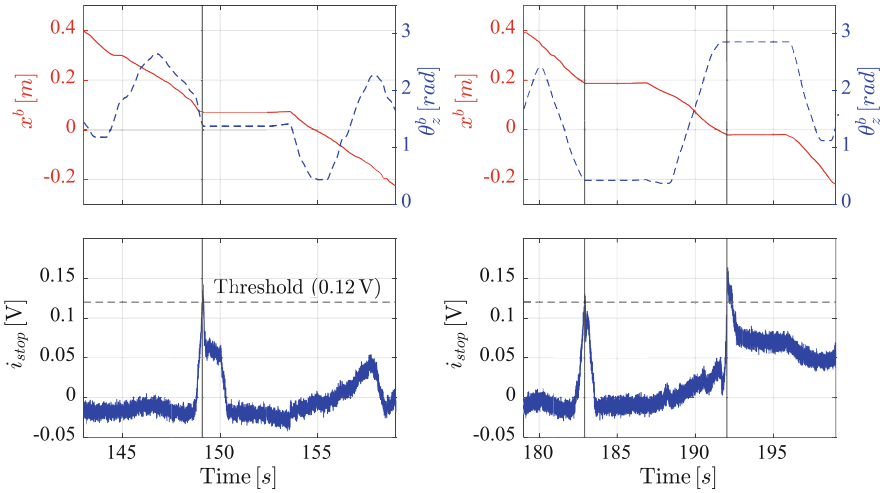


Fig. 7. End effector pose (x^b and θ_z^b) and indicator i_{stop} during the autonomous routing for wires with diameter of 2.5 and 2.0 mm from left to right, respectively.

4 Conclusions and Future Works

The paper showed how a single tactile sensor can be exploited for different tasks in a complex application. In the proposed example, the same sensor has been used to estimate the diameter of an electrical wire, to teach the robot a trajectory

using a lead-through programming approach and to detect forces acting on the routed wire during the autonomous execution in order to stop the task.

A current limitation of the presented application is the possibility of moving the robot on a 2D plane only. Hence, as a future direction, a couple of sensorized fingers will be used to extract new tactile-based indicators in order to achieve the possibility of motion in the 3D space.

References

1. Howe, R.: Tactile sensing and control of robotic manipulation. J. Adv. Robot. **8**, 245–261 (1994)
2. De Oliveira, T.E.A., Cretu, A.M., Da Fonseca, V.P., Petriu, E.M.: Touch sensing for humanoid robots. IEEE Instrum. Meas. Mag. **18**, 13–19 (2015)
3. Jamali, N., Sammut, C.: Material classification by tactile sensing using surface textures. In: 2010 IEEE International Conference on Robotics and Automation, pp. 2336–2341 (2010)
4. Bandyopadhyaya, I., Babu, D., Kumar, A., Roychowdhury, J.: Tactile sensing based softness classification using machine learning. In: 2014 IEEE International Advance Computing Conference (IACC), pp. 1231–1236 (2014)
5. Liu, W., et al.: Discrimination of object curvature based on a sparse tactile sensor array. Micromachines **11**(6), 583 (2020)
6. Prado da Fonseca, V., Alves de Oliveira, T.E., Petriu, E.M.: Estimating the orientation of objects from tactile sensing data using machine learning methods and visual frames of reference. Sensors **19**(10), 2285 (2019)
7. Damian, D.D., Newton, T.H., Pfeifer, R., Okamura, A.M.: Artificial tactile sensing of position and slip speed by exploiting geometrical features. IEEE/ASME Trans. Mechatron. **20**(1), 263–274 (2015)
8. Stachowsky, M., Hummel, T., Moussa, M., Abdullah, H.A.: A slip detection and correction strategy for precision robot grasping. IEEE/ASME Trans. Mechatron. **21**(5), 2214–2226 (2016)
9. Costanzo, M.: Control of robotic object pivoting based on tactile sensing. Mechatronics **76**, 102545 (2021)
10. Costanzo, M., De Maria, G., Natale, C.: Two-fingered in-hand object handling based on force/tactile feedback. IEEE Trans. Rob. **36**(1), 157–173 (2020)
11. Yousef, H., Boukallel, M., Althoefer, K.: Tactile sensing for dexterous in-hand manipulation in robotics-a review. Sens. Actuators A **167**(2), 171–187 (2011)
12. Kappassov, Z., Corrales, J.-A., Perdereau, V.: Tactile sensing in dexterous robot hands - review. Robot. Auton. Syst. **74**, 195–220 (2015)
13. Lee, H.: A survey on robot teaching: categorization and brief review. Appl. Mech. Mater. **330**, 648–656 (2013)
14. Villani, V., Pini, F., Leali, F., Secchi, C., Fantuzzi, C.: Survey on human-robot interaction for robot programming in industrial applications. IFAC-PapersOnLine **51**(11), 66–71 (2018)
15. Cirillo, A., Costanzo, M., Laudante, G., Pirozzi, S.: Tactile sensors for parallel grippers: design and characterization. Sensors **21**(5), 2021 (1915)
16. Pirozzi, S., Natale, C.: Tactile-based manipulation of wires for switchgear assembly. IEEE/ASME Trans. Mechatron. **23**(6), 2650–2661 (2018)
17. Cirillo, A., Laudante, G., Pirozzi, S.: Tactile sensor data interpretation for estimation of wire features. Electronics **10**(12), 1458 (2021)

Design of a Wearable Haptic Device to Mediate Affective Touch with a Matrix of Linear Actuators

Nikolas Ferguson[1], Mehmet Ege Cansev[2(✉)], Anany Dwivedi[2], and Philipp Beckerle[2,3]

[1] Department of Mechanical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[2] Chair of Autonomous Systems and Mechatronics, Department of Electrical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
ege.cansev@fau.de
[3] Department of Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Abstract. Since the beginning of humankind, touch has been a fundamental element of emotional communication between individuals. The increasing number and popularity of prosthetic and assistive robots make processing sense of touch essential to express affective touch. Based on the previous work of authors on design criteria to mediate affective touch, this paper introduces a device for generating stroking motion across the forearm to simulate affective touch. The device offers motion execution in over 50 different patterns by encompassing a matrix of linear actuators, which are designed to perform skin indentation and to simulate a stroking sensation. It also allows multiple mode parameters that can be investigated with a variety of stroking sensations. This can potentially reveal insightful results regarding affective touch.

Keywords: Affective touch · Social touch · Tactile stimulation · Human-machine interfaces · Haptic devices

1 Introduction

Over the past decades, the advancements in the field of sensing and control has led to an increased interest in social and assisstive robotics. This has resulted in a numerous applications of human-machine interfaces in prosthetics, exoskeletons and orthoses [1–3]. It is important to note that current assistive robots have limited capability to display social or affective touch, i.e., tactile processing with a hedonic or motivational component [4]. Therefore, they are not able to sense, deal with or interpret this kind of touch [5].

Stroking with a soft brush creates the best response to low-threshold unmyelinated mechanoreceptors, C-tactile afferents, when using medium-range brush velocities which is most pleasant to the participants [6]. In previous research by Ackerley et al. [7], three different materials, soft brush, fur, and sandpaper,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 507–517, 2023.
https://doi.org/10.1007/978-3-031-16281-7_48

were tested on healthy participants by stroking across the forearm skin. Using soft materials and stroking hairy skin rather than glabrous skin were reported to create a more positive and pleasant experience. The authors also noted that the touch from another felt more pleasant than self-touch [7].

There is relatively few research focusing on technical implementations of devices mediating affective touch but their design can be informed by the existing knowledge about human affective perception. Devices from related work with symbolic non-task-oriented tactile communication, were mainly done with either continuous or vibrotactile stimulation. For example, a robotic stimulator with a brush is used to stimulate the forearm continuously to observe the response of skin mechanoreceptors [6]. Huisman et al. [8,9] used vibration motors in apparatuses with a sleeve design. An array of magnets is used by Culbertson et al. [10] as a haptic sketch prototype before ultimately using voice coil motors for vibrotactile stimulation.

As a result of the technical constraints exhibited by the devices that explore human feelings when subjected to affective touch, there are several open research questions:

- How to provide a pleasant and continuous sensation to the human participants?
- Which operation parameters, such as indentation speed, force, type of movement, as well as movement speed will be relevant in the design process?
- Do different indentation patterns arouse different emotional responses by the participants?

The end goal of this work is to answer the above mentioned research questions. In this work, we take the first step towards achieving the end goal with the proposed wearable haptic device. More specifically, in this work we design, program, build and functionally test a proof-of-concept haptic stimulation device to evoke a feeling of affective touch in the participants. Once attached, the device should be able to create indentation on the participant's skin to have a pleasant feeling from the device. With the proposed design, we achieved different stimulation parameters and patterns. The effects of these will be validated in a systematic manner in a user study which is being designed.

To this end, Sect. 2 introduces the fundamental test design for user studies and an overview of the device concept. In Sect. 3, detailed information regarding psychological and biomechanical requirements are given, component selection is explained, and finally the complete designed is presented. Section 4 gives a description of the control interface, different stimulation modes and pattern variations. Ultimately, a conclusion and outlook on future work are given in Sect. 5.

2 Concept

To understand how to evaluate the effects of affective touch in human-machine interaction applications, we present a system concept and propose a test design that mediate affective touch to elicit emotional responses.

2.1 System Concept

The system based on two main elements: i) the wearable device, which comprises of actuators and sensors and ii) a control interface, that controls the device. The concept of the device can be seen in Fig. 1.

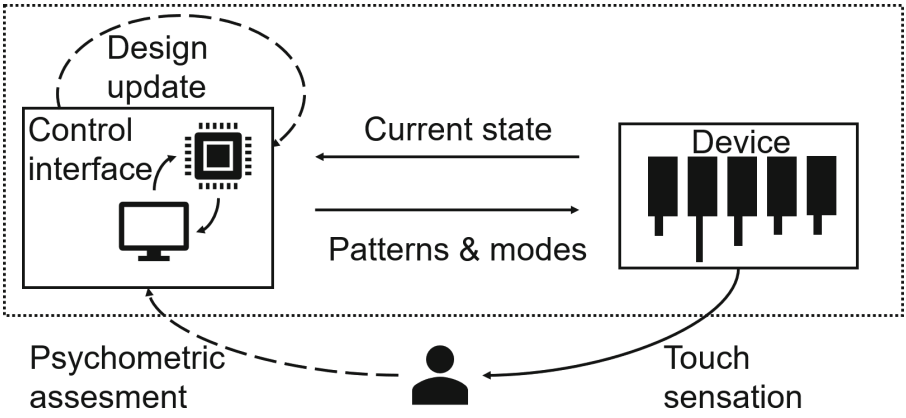


Fig. 1. The concept of the proposed device to emulate affective touch of human-machine interaction. The dotted box represents the system boundary. The dashed arrows represent offline (after testing) data acquisition and processing. The user receives touch sensation from the device and provides a psychometric assessment to the system.

The control interface of the device , Arduino Mega [11], sends control commands in the form of pattern and modes. For a closed-loop control, the device sends its current state, e.g., actuator positions, back to the control interface. Additionally, forces exerted on the skin are only monitored for now and planned to be closed-loop controlled. The actuators are utilized to administer affective touch via skin indentation. The device offers a flexible range of operation for various tactile stimulation parameters that play a critical role to mediate affective touch, such as, lateral stimulation velocity and indentation force on the skin [5]. Beyond these parameters, different stimulation patterns and propagation velocities for skin indentation, i.e., lateral velocity of perceived touch, along the forearm are planned to be investigated as the independent variables in a future study. Based on the psychometric assessment, the design and stimulation patterns and modes are going to be updated to achieve the most pleasant and continuous touch sensation.

2.2 Planned Fundamental Test Design

To analyze the effects of different stimulation patterns and parameters, we propose a user study design. The device is attached to the forearm of the participants

while they sit comfortably on a chair while resting their hand on the table. Participants wear headphones to not get distracted by the sounds of the device. Furthermore, their visual contact with the device is disconnected by covering their forearm to minimize the effects of visual stimuli, facilitating a realistic experience of affective touch.

The study focuses on the evaluation of psychological aspects of affective touch rather than the biological reaction of human skin, e.g., firing rate of the fibers responsible for affective touch. After the test sequence concludes, a psychometric assessment can be utilized to implement design updates to the control interface, to improve the whole system. As frequently evaluated in the previous works [6,8,12,13], pleasantness and, as an inherent requirement for discrete stimulation, continuity are planned to be evaluated by the participants through Likert scales [14]. In addition to psychometric measures, psychophysiological assessment techniques, such as electrodermal activity (EDA) [15], can also be used to quantify the pleasantness of participants.

3 Design

To the best of our knowledge, previous studies have mainly focused on voice coil motors for vibrotactile stimulation [8–10]. However, we expect that using linear actuators will create a more realistic feeling when trying to emulate affective touch on the human forearm. Therefore, a matrix of linear actuators was chosen to simulate the realistic feeling of touch.

3.1 Psychological and Biomechanical Requirements

Some significant requirement set for the device are: i) indentation forces, ii) lateral speed, iii) indentation depth, and iv) material hardness of the contact point. From previous research, it was concluded that lateral velocities from 1 cm/s up to 13.5 cm/s are perceived as pleasant [6,8,10,12,13,16,17]. However, corresponding skin receptors, C-tactile afferents, do not respond strongly to the tactile stimulations with higher velocities [13]. Although C-tactile afferents are activated at 5 mN [18], research has shown that stronger indentation forces of 0.1 N - 2 N [6,10,19] are effective and achieve responsiveness. Furthermore, a displacement of 1.5 mm normal to the skin surface is sufficient for stimulation [20] and using soft materials is suggested to perform indentation [12,16].

3.2 Selection of Actuators and Sensors

Actuators and sensors are selected based on the psychological and biomechanical requirements. Unlike commonly used actuators in haptic devices such as, vibration motors or microfluidic actuators, stroking motion by linear actuators with proper dimensions and performance could resemble fingerlike motion due to the differences in movement modalities. After examining ten different linear actuators, the PQ12-P, Actuonix, Saanichton [21] was the chosen actuator.

Table 1. Comparison of affective touch requirements and selected actuator specifications

Affective touch requirements		Actuonix PQ12 30:1 specifications [21]	
Indentation force	0.1–2 N[a]	Max. force (lifted) Max. efficiency point	18 N 8 N
Indentation length	1.5 mm [20]	Stroke length Positional repeatability	20 mm ± 0.1 mm
Indentation velocity	–	Max. stroking velocity Max. efficiency point	28 mm/s 20 mm/s
Caressing velocity	1–13.5 cm/s[b]	Stroke propagation velocity	Pulse width and delay controlled

[a] The range of indentation force is obtained by combining the findings of [6,10,19].
[b] The range of caressing velocity is obtained by combining the findings of [6,8,10,12,13,16,17].

Unlike other options, the selected actuator fulfills force and stroke requirements by using simple control approaches as well as providing a large range of stroke velocity with its lightweight, compact, adequately small and reliable design.

When performing skin indentation to mediate affective touch, the maximum force needed is under 18 N [6,18,19]. Although there are other models of the selected actuator with higher reduction ratio, the one with smallest reduction ratio is selected as it satisfies the force requirements and offers the highest stroking speed, i.e., 28 mm/s, as shown in Table 1. This crucially allows to experiment with a greater range of indentation velocity which is an overlooked stimulation parameter by the previous research. The maximum stroke length of the chosen actuator is 20 mm, which satisfies the requirement for indentation length abundantly with a precision of 0.1 mm as the required indentation depth is 1.5 mm and the actuators are placed just above the skin. As for the controller options, "S" stands for limit switch, "P" stands for potentiometer position feedback and "R" stands for RC linear servo. The "P" actuator model was chosen in order to control the stroke length so that the indentation depth is controlled [21]. The sensor selection comprised of using force sensitive resistors (FSR's) as pressure sensors. A program was designed to detect force values between the actuator contact point and the skin of the user. Although the force sensors are only used to monitor the force applied to the skin currently, the control algorithm is planned to be updated to cascaded force and position control.

3.3 Mechanical Design

As a means to incorporate the actuators, sensors, control device and connectors, a mechanical structure is needed to blend these single components into an actual wearable device on the forearm.

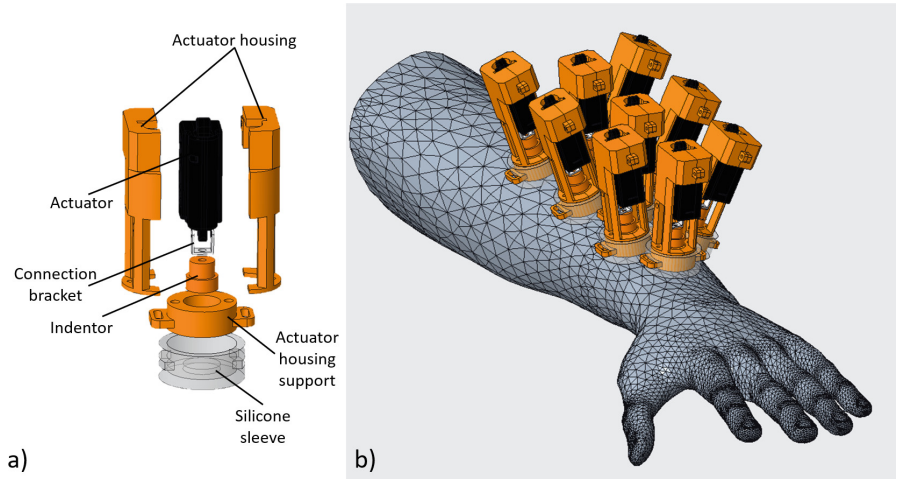


Fig. 2. Mechanical design of the device. Subfigure a) shows an exploded view of one actuator unit while subfigure b) shows a concept of the proposed system with nine actuators mounted on the forearm.

Figure 2a shows the exploded view of a single actuation unit. Each actuation unit is composed of three main component blocks. i) The first block comprises of the actuator and its housing and the connection bracket, which is responsible to connect the end-effector to the actuator. ii) The second block is the indentor, that transmits the forces from the actuator to the human skin over a larger surface area than the end-effector of the actuator. iii) Finally, the third block includes the actuator housing support, that serves as the final connection piece between the actuator housing and the human forearm, and a silicone sleeve, which is a padded piece to create a pleasant feeling on the forearm, instead of having hard and uncomfortable 3D-printed plastic parts touching the skin.

Nine of the single actuator units are placed on the forearm as shown in Fig. 2b. Together they form a matrix of actuators across the top side of the forearm. As having all nine actuators in a single line across the forearm would not allow for many different types of stimulation patterns and would not fit on most forearms, it was decided to have two actuators side by side in every second row. Rubber cords shown in Fig. 3 are used to fasten the actuators to the arm. This gives room for adjustable placement of the units to be adaptable for different arm sizes. The current status of the device intended for the scientific experiments is shown in Fig. 3. Although the device appears bulky, the main use case for the device for the moment is in a laboratory setting to gain insights regarding affective touch. For such applications the size of the device should not effect its usability. By eliciting emotional responses to improve the pleasantness and continuity of touch sensation, we foresee that affective touch can play an important role in the embodiment of prosthetic devices or increasing the quality of user experience during interaction with assistive robotics. Therefore, bulkiness

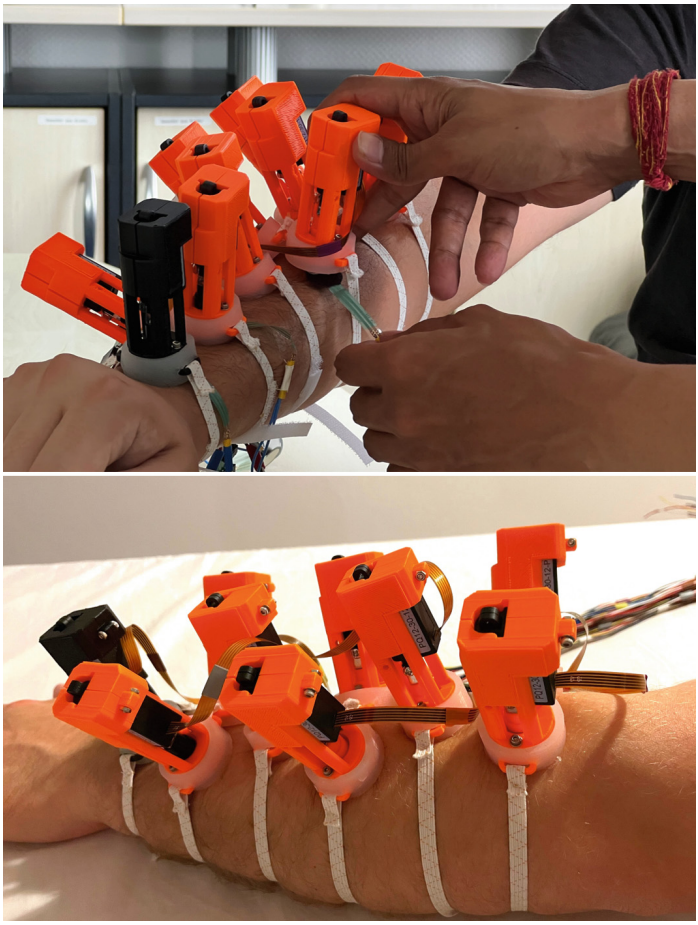


Fig. 3. The developed device worn by a participant on their forearm. The top subfigure shows the experimenter attaching an FSR sensor to detect the forces exerted by the sensor. The bottom subfigure shows an instance of the worn device where the setup without FSRs is complete.

is a long-term problem that needs to be solved to improve wearability, and hence, inclusion in everyday use.

4 Control Interface

A control interface is designed to send and receive data to and from the user via sensors and actuators. While the low-level control regulates the indentation length of the actuators, the focus of this section is on the high-level control of the device which consists of stimulation patterns and modes.

4.1 Patterns

Operating multiple actuators in distinct combinations allows for the different stimulation patterns which refer to the order in which the actuators are performing skin indentation. As each actuator requires digital and analog pins to control the velocity, position, and direction of the end-effector, an Arduino Mega [11] alongside H-bridges is used to drive the actuators.

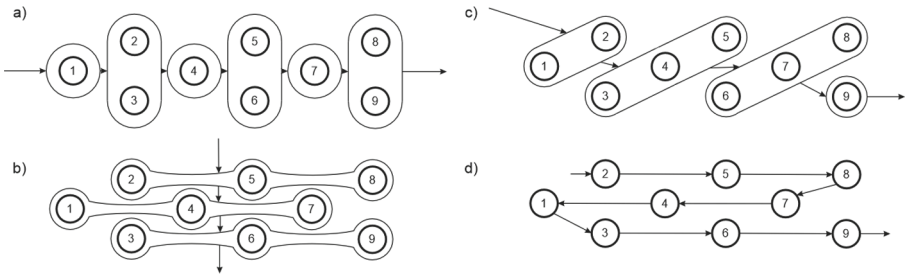


Fig. 4. Stimulation patterns

Figure 4 depicts possible stimulation patterns for the device by assigning numbers 1–9 to the actuators in the matrix structure with triangulated rows. This allows for more stimulation pattern possibilities. Pattern a) shows a horizontal pattern approach. The sequence begins with actuator 1 and ends with actuator 8 and actuator 9. Then the sequence can be either repeated, or combined with another pattern and is a typical stroking pattern for when a human strokes the arm of another human being. Pattern b) and c) implement similar sequences in vertical and diagonal directions, respectively. Pattern d) is a snake-like pattern that begins with the top left actuator and ends with the bottom right one. Beyond these patterns, spatial flexibility offers plenty different alternatives including a random pattern.

The aforementioned patterns are defined over spatial dimensions without including temporal dimensions. Although the order of stimulation does not change, varying the stroke duration of the actuators, i.e., flexibility in temporal dimensions, can lead to a totally different perception of touch. Considering each actuator unit as a finger makes it easier to appreciate the intuitiveness and necessity of temporal dimensions in the patterns. As a result, there are over 50 possibilities to form patterns which allows researchers to explore the response of skin receptors to multitudinous stimulation patterns.

4.2 Modes

All actuation patterns can have different modes. Modes contain different parameters such as stroking velocity, pattern propagation speed, indentation depth or

indentation force. By changing the pulse width and delay parameters in the algorithm, one can easily examine the effects of different propagation velocities of stroking motion on pleasantness and continuity of the touch sensation. Although it is not easy to precisely determine the indentation force and length with compliant behavior due to silicone sleeve at the contact point and rubber cords around the arm, the device provides ranges of 0.2 N to 20 N and 0 mm to 5 mm for indentation force and length, respectively, which contain the affective touch requirements with a pleasant sensation. As the indentation force is controlled indirectly through the stroke length of actuators, PWM signals offers adequate resolution to test the different indentation force values that are reported to be mediating affective touch. Since the previous devices that mediate affective touch uses mostly either continuous lateral movements, e.g., via a brush, or vibration motors, the effects of vertical stroking velocities on pleasantness have not been investigated so far. With flexible indentation velocities up to 28 mm/s provided by our design, we aim to evaluate those effects at multiple velocities.To have multiple modes and patterns combined, creates a very large variety of possibilities for further work.

5 Conclusion

In this paper, we propose a wearable haptic device that mediates affective touch to improve the quality of human-machine interaction. Unlike previously developed devices, we proposed a matrix of linear actuators for discretization of tactile stimuli. Such a design approach allows testing various stimulation patterns and modes at the expense of potential reduction of continuity feeling. Selecting linear actuators instead of vibration motors aims to improve the realism of tactile perception. Moreover, when the temporal flexibility is considered, the matrix of actuators vastly enhances the range of stimulation patterns compared to previous works by adding the second spatial dimension to the array of actuators. Furthermore, the flexibility in various modes helps determining the optimal operation range for stimulation parameters to mediate affective touch. The device was tested on a participants with various operation modes and patterns as part of a technical pre-assessment, without statistical analysis, verifying all the functions of the device. To validate the effectiveness of the developed device in a systematic manner, user studies are being designed, focusing on psychometric and psychophysiological assessments. This will allow evaluation of the psychological factors that are highly dependent on affective touch with the changing stimulation conditions.

To further develop the design, we consider a fixed structure that all the linear actuators are attached to facilitate a better repeatability for the actuator placement. A fixed structure is avoided at this stage, as it requires meticulous positioning of actuators to achieve ideal contact with human skin for different participants. Although the silicone layer introduces little flexibility at the point of contact, a mechanism based on remote center of compliance (RCC) can be implemented so that the ideal contact can be guaranteed. Additionally, measuring or estimating lateral forces along the skin might provide insight to utilize

affective touch to create stronger emotional responses in human-machine interaction. With the suggested improvements, the device has the potential to enlighten the undiscovered effects of affective touch in human-robot interaction.

Acknowledgments. This work received support from the Mercator Research Center Ruhr (MERCUR) (Grant Number: An-2019-0032). The support by the Mercator Research Center Ruhr is acknowledged.

References

1. Dollar, A.M., Herr, H.: Lower extremity exoskeletons and active orthoses: challenges and state-of-the-art. IEEE Trans. Rob. **24**(1), 144–158 (2008)
2. Beckerle, P., et al.: A human-robot interaction perspective on assistive and rehabilitation robotics. Front. Neurorobot. **11**, 24 (2017)
3. Beckerle, P., et al.: Feel-good robotics: requirements on touch for embodiment in assistive robotics. Front. Neurorobot. **12**, 84 (2018)
4. Morrison, I.: Ale meta-analysis reveals dissociable networks for affective and discriminative aspects of touch. Hum. Brain Mapp. **37**(4), 1308–1320 (2016)
5. Ege Cansev, M., Nordheimer, D., Andrea Kirchner, E., Beckerle, P.: Feel-good requirements: neurophysiological and psychological design criteria of affective touch for (assistive) robots. Front. Neurorobot. **15**, 661207 (2021)
6. Löken, L.S., Wessberg, J., Morrison, I., McGlone, F., Olausson, H.: Coding of pleasant touch by unmyelinated afferents in humans. Nat. Neurosci. **12**(5), 547–548 (2009)
7. Ackerley, R., Saar, K., McGlone, F., Backlund Wasling, H.: Quantifying the sensory and emotional perception of touch: differences between glabrous and hairy skin. Front. Behav. Neurosci. **8**, 34 (2014)
8. Huisman, G., Frederiks, A.D., van Erp, J.B.F., Heylen, D.K.J.: Simulating affective touch: using a vibrotactile array to generate pleasant stroking sensations. In: Bello, F., Kajimoto, H., Visell, Y. (eds.) EuroHaptics 2016. LNCS, vol. 9775, pp. 240–250. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42324-1_24
9. Huisman, G., Darriba Frederiks, A., van Dijk, B., Hevlen, D., Krose, B.: The TaSST: tactile sleeve for social touch. In: 2013 World Haptics Conference (WHC), 14–17 April 2013, pp. 211–216. IEEE (2013)
10. Culbertson, H., Nunez, C.M., Israr, A., Lau, F., Abnousi, F., Okamura, A.M.: A social haptic device to create continuous lateral motion using sequential normal indentation. In: IEEE Haptics Symposium (HAPTICS), vol. 2018, pp. 32–39. IEEE (2018)
11. Arduino: Arduino Mega 2560 Rev3 datasheet: specifications. https://arduino.cc
12. van Stralen, H.E., van Zandvoort, M.J.E., Hoppenbrouwers, S.S., Vissers, L.M.G., Kappelle, L.J., Dijkerman, H.C.: Affective touch modulates the rubber hand illusion. Cognition **131**(1), 147–158 (2014)
13. Crucianelli, L., Krahé, C., Jenkinson, P.M., Fotopoulou, A.K.: Interoceptive ingredients of body ownership: affective touch and cardiac awareness in the rubber hand illusion. Cortex **104**, 180–192 (2018)
14. Likert, R.: A technique for the measurement of attitudes. Arch. Psychol. **140**, 5–55 (1932)
15. Shukla, J., Barreda-Angeles, M., Oliver, J., Nandi, G., Puig, D.: Feature extraction and selection for emotion recognition from electrodermal activity. IEEE Trans. Affect. Comput. **12**(4), 857–869 (2019)

16. Essick, G.K., James, A., McGlone, F.P.: Psychophysical assessment of the affective components of non-painful touch. NeuroReport **10**(10), 2083–2087 (1999)
17. Crucianelli, L., Metcalf, N.K., Fotopoulou, A.K., Jenkinson, P.M.: Bodily pleasure matters: velocity of touch modulates body ownership during the rubber hand illusion. Front. Psychol. **4**, 703 (2013)
18. Vallbo, A.B., Olausson, H., Wessberg, J.: Unmyelinated afferents constitute a second system coding tactile stimuli of the human hairy skin. J. Neurophysiol. **81**(6), 2753–2763 (1999)
19. Nordin, M.: Low-threshold mechanoreceptive and nociceptive units with unmyelinated (C) fibres in the human supraorbital nerve. J. Physiol. **426**, 229–240 (1990)
20. Biggs, J., Srinivasan, M.: Tangential versus normal displacements of skin: relative effectiveness for producing tactile sensations. In: Proceedings 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, HAPTICS 2002, pp. 121–128 (2002)
21. Actuonix Motion Devices Inc.: PQ-12 datasheet: Specifications. Victoria BC (2016). www.actuonix.com

Motor-Unit Ordering of Blindly-Separated Surface-EMG Signals for Gesture Recognition

Mattia Orlandi[1(✉)], Marcello Zanghieri[1], Davide Schiavone[2], Elisa Donati[3], Francesco Conti[1], and Simone Benatti[4]

[1] University of Bologna, Bologna, Italy
mattia.orlandi@studio.unibo.it
[2] Swiss Federal Institute of Technology Lausanne, Lausanne, Switzerland
[3] University of Zurich, Zurich, Switzerland
[4] University of Modena e Reggio Emilia, Reggio Emilia, Italy

Abstract. Hand gestures are one of the most natural and expressive way for humans to convey information, and thus hand gesture recognition has become a research hotspot in the human-machine interface (HMI) field. In particular, biological signals such as surface electromyography (sEMG) can be used to recognize hand gestures to implement intuitive control systems, but the decoding from the sEMG signal to actual control signals is non-trivial. Blind source separation (BSS)-based methods, such as convolutive independent component analysis (ICA), can be used to decompose the sEMG signal into its fundamental elements, the motor unit action potential trains (MUAPTs), which can then be processed with a classifier to predict hand gestures. However, ICA does not guarantee a consistent ordering of the extracted motor units (MUs), which poses a problem when considering multiple recording sessions and subjects; therefore, in this work we propose and validate three approaches to address this variability: two ordering criteria based on firing rate and negative entropy, and a re-calibration procedure, which allows the decomposition model to retain information about previous recording sessions when decomposing new data. In particular, we show that re-calibration is the most robust approach, yielding an accuracy up to 99.4%, and always greater than 85% across all the different scenarios that we tested. These results prove that our proposed system, which we publish open-source and which is based on biologically plausible features rather than on data-driven, black-box models, is capable of robust generalization.

Keywords: sEMG · Blind source separation · Gesture classification

1 Introduction

Nowadays, hand gesture recognition is gaining traction in the field of human-machine interfaces (HMIs), and has shown great potential in several contexts such as augmented reality, robotics and prosthetics [1]. One of the most

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 518–529, 2023.
https://doi.org/10.1007/978-3-031-16281-7_49

versatile sources of signal to convey information about hand gestures is the electromyography (EMG) signal [2–4], which describes the electrical activity of the muscle. Specifically, surface electromyography (sEMG) is particularly suited for wearable myoelectric control systems (MCSs), because it allows to collect the physiological bio-potential directly from the skin via surface electrodes, granting a fully non-invasive setup [5].

Typically, the control systems that recognize hand gestures acquire the sEMG signal and process it by mapping the signal patterns onto a given set of gestures. The mapping operation is usually performed using either hand-crafted features—such as root mean square (RMS), mean absolute value (MAV), and discrete wavelet transform (DWT) [6]—that can then be used to train a classifier, or data-driven, deep learning (DL)-based feature learning such as convolutional neural networks (CNNs) [7] or temporal convolutional networks (TCNs) [8].

A more bio-inspired approach is to decompose the sEMG signal into its fundamental elements [9], the motor unit action potential trains (MUAPTs), which can be achieved via blind source separation (BSS): in fact, the problem of recovering MUAPTs from the EMG recordings can be restated in terms of the typical cocktail party problem, usually addressed using BSS techniques. Such techniques are indeed able to reconstruct the MUAPTs, and the firing of these motoneurons (MNs) provides a major control information on strength and duration of muscle contractions, constituting a bio-inspired feature in the gesture recognition framework.

The most adopted BSS algorithms for spike extraction are independent component analysis (ICA) [10] and convolution kernel compensation (CKC) [11]. Validation of these approaches is done on synthetic datasets, in which sEMG signals are generated using models of MNs, tissues, and sensor positions, derived from physiological information.

A major limitation of the current BSS methods is that, by their nature, such algorithms return reconstructed sources without any particular ordering [10], making their use difficult for interfaces based on classical recognition frameworks, which must work reliably even in different recording sessions.

In this paper, we propose two methods for consistently ordering source MUAPTs. The first method, of a more theoretical/statistical nature, is based on negative entropy [10], whilst the second is more empirical/physiological and is based on the firing rate of motor units (MUs). These ordering methods make it possible to apply the BSS as a feature extraction in different recording sessions, with the advantages of a physiological information which we leverage for performing gesture recognition. The main contributions of this work are: (1) presentation of negative-entropy-based and firing-rate-based methods for ICA BSS, validated on synthetic data, (2) end-to-end framework based on neural spikes for gesture classification, including a re-calibration strategy for new sessions, (3) testing on real sEMG data, yielding up to 99.4% accuracy in a 1-DoF setting targeting inter-session accuracy, the SoA metric for long-term generalization of sEMG recognition systems [12,13]. In addition, we publish open-source all the code used to reproduce our work at https://github.com/nihil21/semg-bss.

2 Materials And Methods

2.1 Data Model

The central nervous system conveys commands to the muscles by means of sequences of electric impulses via alpha-motoneurons (α-MNs), whose *somas* are located in the spinal cord and whose terminal axons innervate a group of muscle fibers. A MU consists of an α-MN and the muscle fibers it innervates [14]. The electrical activity of a MU, the motor unit action potentials (MUAPs), can be detected by electrodes, either placed inside the target muscle or on the skin above the target muscle: the former setting results in intramuscular electromyography (iEMG) recordings, the latter in surface electromyography (sEMG) recordings. In particular, sEMG is preferred since it is painless and non-invasive [14]. Nevertheless, sEMG signals are a temporal and spatial summation of several MUAPTs resulting in a complex inteference pattern [14], which can be modelled as follows [11,15]:

$$x_i(t) = \sum_{j=1}^{N} \sum_{\tau=0}^{L-1} h_{ij}(\tau) s_j(t-\tau) + \omega_i(t), \quad i = 1, \ldots, M, \quad t = 0, \ldots, D_R \tag{1}$$

where (1) $x_i(t)$ is the i-th EMG channel, (2) $h_{ij}(\tau)$ is the action potential of the j-th MU as recorded at the i-th channel, (3) $s_j(t) = \sum_k \delta\left(t - T_j^{(\text{discharge})}(k)\right)$ is the MUAPT of the j-th MU, modelled as a sum of delta functions where $T_j^{(\text{discharge})}(k)$ is the timing of the k-th discharge, (4) $\omega_i(t)$ is the additive noise at the i-th channel, (5) L is the duration of MUAPTs, (6) N is the number of MUs, (7) M is the number of channels, and (8) D_R is the duration of the signal measured in discrete time samples. In particular, h_{ij} governs both the instantaneous mixing process between the i-th and j-th sources and the temporal mixing process, as it acts as a finite impulse response (FIR) filter with length L [11,15]. For each $\tau = 0, \ldots, L-1$, h_{ij} are assumed to be constant for the duration of the recording D_R [15]. The aim is to recover as many sources $s_j(t)$ as possible from the recorded signals $x_i(t)$, assuming that the identified sources are either independent or sparse. The non-identified MUs are modelled into the additive noise term in Eq. 1 [11,15].

A convolutive mixture with FIR filters such as the one in Eq. 1 can always be represented as a linear and instantaneous mixture of an extended vector of sources with shape $NL \times 1$, which includes the original sources and their $L-1$ delayed replicas [11,15–17]. In order to increase the conditionality of the mixing process (i.e., the ratio between the number of observations and the number of sources), the vector of observations is extended by adding $f_e - 1$ delayed replicas, where f_e is a given extension factor. The extended model is defined as:

$$\bar{\mathbf{x}}(t) = \bar{\mathbf{H}}\bar{\mathbf{s}}(t) + \bar{\boldsymbol{\omega}}(t) \tag{2}$$

where $\bar{\mathbf{x}}(t)$, $\bar{\mathbf{s}}(t)$, $\bar{\boldsymbol{\omega}}(t)$ are the extended vectors of observations, sources and noise, respectively, while $\bar{\mathbf{H}}$ is the instantaneous mixing matrix. Assuming $\bar{\mathbf{H}}$ is full-

rank, in order to address the inverse problem the system in Eq. 2 must be overdetermined, namely the number of extended observations must be higher than the number of sources multiplied by the length of the filters: $Mf_e \geq N(L+f_e-1)$ [11]. In general, it was observed that values of f_e greater than 16 did not substantially influenced the results of the decomposition [15].

2.2 ICA-Based Algorithm

The proposed system relies on convolutive ICA algorithms, and in particular on Hyvarinen and Oja's fast fixed-point algorithm for ICA (FastICA) [10]. In fact, ICA can be used as a preprocessing tool to decompose EMG signals into their constitutive MUAPTs [14–19]. The ICA model presents some requirements [14] such as (1) a linear and instantaneous mixing process, and (2) independence of the source signals : thanks to the extension process in Eq. 2 the former requirement is fulfilled, whereas the latter is not, since the delayed replicas are not independent. However, as explained below, the FastICA algorithm makes use of contrast functions for source separation: these functions were designed to maximize the non-Gaussianity and, only indirectly, the independence of the estimated sources [10]. MUAPTs are leptokurtic [14], namely their probability distribution has heavier tails compared to a normal distribution, and thus they tend to be more sparse [15]. Therefore, the contrast functions can be used as a measure of sparseness rather than a measure of independence, and it was observed that FastICA still manages to recover the extended sources even if they are not independent [15].

The first steps of FastICA, common to many general-purpose BSS methods, are centering and whitening of the extended vector of observations $\bar{\mathbf{x}}(t)$: after subtracting its mean, the vector is linearly transformed into a new vector $\hat{\mathbf{x}}(t)$ which is "white", namely its covariance matrix is the identity matrix. Whitening reduces the number of parameters to be estimated [10], and it can be performed by means of the ZCA method: $\hat{\mathbf{x}}(t) = \mathbf{W}\bar{\mathbf{x}}(t)$ with $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}^{-1/2}\mathbf{V}^{\intercal}$, where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\intercal}$ is the singular value decomposition (SVD) of the covariance matrix $\mathbf{K}_{\bar{\mathbf{x}}\bar{\mathbf{x}}} = \mathbb{E}[\bar{\mathbf{x}}(t)\bar{\mathbf{x}}^{\intercal}(t)]$. The procedure of signal extension and whitening is referred to as "convolutive sphering" [16]. In particular, as per [15], only the eigenvalues greater than a certain threshold are considered, and such threshold is set to the average of the smallest half of the eigenvalues: this regularization procedure should help reducing the effect of noise.

The FastICA procedure is based on a fast-fixed point iteration scheme for maximizing the non-Gaussianity of the estimated sources [10]. The non-Gaussianity of a random variable y is measured by means of negative entropy (or neg-entropy) $J(y)$; since computing neg-entropy exactly would require a possibly non-parametric estimate of the probability density function of the estimated sources, an approximation based on the maximum-entropy principle [10] is used in practice:

$$J(y) \propto \left(\mathbb{E}[G(y)] - \mathbb{E}[G(\nu)]\right)^2 \tag{3}$$

where ν is a random Gaussian variable with zero mean and unit variance, and G is a contrast function [10]. A common choice for the contrast function is

$G(u) = \log(\cosh(u))$, since it is more robust to outliers and signal artifacts [15]; additionally, to prevent the algorithm from converging to the same source multiple times, the Gram-Schmidt orthogonalization step is performed [10,15]. A fast convergence of FastICA requires a suitable initialization point of the projection vector: therefore, peaks in the squared summation of the whitened observations are detected, and each separation vector is initialized to the value of a peak, chosen randomly among the 25% highest peaks, as per [15,18].

In order to reduce the variability in the discharge timings of the estimated MUs, a second iterative procedure is employed: spike trains are estimated by performing peak detection on the squared source and by classifying the detected peaks with K-means with $k = 2$ to discriminate between actual spikes and noise. This refinement procedure is repeated as long as the coefficient of variation (CoV) of the inter-spike interval (ISI) of the estimated MUs decreases, as per [15,17,18].

A silhouette measure (SIL), defined as the normalized difference between the sum of point-to-centroid distances within cluster and between clusters [15,17,20], is computed on the estimated spike train, and the corresponding separation vector is accepted only if SIL is greater than a given threshold `th_sil`. The process is repeated M_t times, where M_t is a hyper-parameter indicating the number of target MUs (i.e., the number of MUs the system tries to extract): in the end, the separation matrix $\mathbf{B}$ will contain at most M_t rows, namely the separation vectors for at most M_t source MUAPTs.

Finally, a post-processing step is applied in order to remove either inactive MUs, namely MUs with a firing rate less than a given threshold (in spike/s) [19,20], and duplicates: in fact, as mentioned before, the extension step in Eq. 2 introduces delayed replicas of the sources, and thus the decomposition algorithm may converge to such duplicates, despite the orthogonalization step. In particular, delayed replicas are detected by checking those MUAPTs that share more than 50% firing events within a ± 1 ms tolerance window, and only the MU with the highest SIL is kept [19,20].

The procedure described so far is designed to be used for calibrating the decomposition model at the beginning of the recording session: then, relevant parameters such as the mean vector $\boldsymbol{\mu}_{\bar{\mathbf{x}}}$, the whitening matrix $\mathbf{W}$, the separation matrix $\mathbf{B}$ and the thresholds for spike/noise classification $\boldsymbol{\Gamma}$ are stored, and can be used in an online fashion to efficiently decompose raw sEMG signals during the same session [17,18].

Additionally, the model is re-calibrated on sEMG data from a new recording session. First of all, the mean vector $\boldsymbol{\mu}_{\bar{\mathbf{x}}}$ and the whitening matrix $\mathbf{W}$ are updated via the following momentum-controlled rule:

$$\boldsymbol{\mu}_{\bar{\mathbf{x}}} \leftarrow (1-\beta) \cdot \boldsymbol{\mu}_{\bar{\mathbf{x}}} + \beta \cdot \boldsymbol{\mu}_{\bar{\mathbf{x}}}^{(\text{new data})}, \quad \mathbf{W} \leftarrow (1-\beta) \cdot \mathbf{W} + \beta \cdot \mathbf{W}^{(\text{new data})} \tag{4}$$

where β is momentum, and *new data* means parameters fit on the new session only. Then, assuming that in the previous session P MUs ($P < M_t$) were extracted (i.e., the separation matrix $\mathbf{B}$ has P rows), the model estimates up to $(M_t - P)$ new MUs from the current session; the corresponding separation vectors are used to update the separation matrix $\mathbf{B}$ by vertically concatenating them to

B itself (i.e., the new rows represent the coefficients of the new extracted MUs). The orthogonalization step prevents the estimation of separation vectors already identified during the previous session (i.e., already present in **B**), and for this reason the model may estimate fewer MUs than M_t. If Q MUs are extracted in total ($P \leq Q \leq M_t$), the separation matrix **B** has Q rows, and the vector of spike/noise thresholds $\mathbf{\Gamma}^{(\text{new data})}$ has Q entries. Finally, the first P entries of the vector of spike/noise thresholds $\mathbf{\Gamma}$ are updated with the same momentum-controlled rule as $\boldsymbol{\mu}_{\bar{\mathbf{x}}}$ and **W**:

$$\begin{aligned} \Gamma_i &\leftarrow (1-\beta)\cdot\Gamma_i + \beta\cdot\Gamma_i^{(\text{new data})}, \quad i = 1,\ldots,P \\ \Gamma_j &\leftarrow \Gamma_j^{(\text{new data})}, \quad j = P+1,\ldots,Q \end{aligned} \tag{5}$$

The motivation for using momentum is to obtain values that are a good fit for both the old data and the current session's data. In this work, we use $\beta = 0.5$.

3 Experimental Results

The proposed system was validated through both a synthetic sEMG dataset with known ground truth MUAPTs and real sEMG signals recorded during the performance of a set of gestures, for which a ground truth label is provided. As per [15,17,18,21], we set (1) the maximum number of iterations to 100, (2) the threshold for convergence of FastICA and source improvement iterations to 10^{-4}, (3) the threshold for SIL to 0.9, (4) the threshold for considering a MU as active to 8 spike/s.

3.1 Decomposition Validation

To assess the decomposition accuracy of the system a synthetic dataset was used. The dataset was provided by Mohebian et al.[22]: it was generated using a planar volume conductor model, and consists of 15 simulated sEMG signals with 90 channels, a length of 16 s and a sampling frequency of 4096 Hz. The muscle excitation for each simulated signal was set to either 10%, 30% or 50% Maximum Voluntary Contraction (MVC).

As per [22], a MU was considered correctly identified when at least 30% of its firings were time-locked in a window of ±0.5 ms with a ground truth MU. For each identified MU, the accuracy of the decomposition was assessed by computing the following metrics [22]:

- rate of agreement (RoA): $\frac{\text{TP}}{\text{TP+FN+FP}}$
- precision: $\frac{\text{TP}}{\text{TP+FP}}$
- recall: $\frac{\text{TP}}{\text{TP+FN}}$

As per [15], we set the extension factor $f_e = 16$, while we varied the number of target MUs M_t depending on the value of the MVC. Results are shown in Table 1: "#MUs GT" is the number of active ground-truth MUs, "#MUs extracted" is the

actual number of MUs the system manages to extract, and "#MUs (identified)" is the number of extracted MUs correlating with a ground truth MU based on the time-locking criteria explained before. As it can be seen, our system is able to detect only a small portion of the ground-truth MUs (the ones simulated near the locations of the electrodes): this is expected, since MUs far from the electrodes contribute to the physiological noise; as a comparison, the modified gradient convolution kernel compensation (gCKC) extracted on average 16.41 ± 4.18 MUs [22]. Noteworthily, the identified MUs are very strongly correlated with the ground truth: in particular, the larger the MVC, the fewer the extracted and identified MUs are, whereas the correlation metrics remain firmly high. Differently from [22], in which the authors added Gaussian noise with a given signal-to-noise ratio (SNR) to the simulated sEMG signals, these results were obtained without considering any additional noise, apart from the simulated physiological noise and the contribution of unidentified distant MUs.

3.2 Gesture Classification

Concerning the gesture classification task, the proposed system was tested on the putEMG dataset, collected by Kaczmarek et al. [23] using a 24-electrode matrix placed on the forearm. It includes recordings sampled at 5120 Hz from 44 healthy subjects and from two sessions, separated by at least one week. Each recording contains repetitions of 7 active gestures separated by an idle state. For each subject and session, the dataset provides three signals:

Table 1. Number of extracted and identified MUs, RoA, precision and recall, averaged over 5 simulated signals for each MVC value: results are reported as mean ± std.

MVC	M_t	#MUs GT	#MUs extracted	#MUs identified	RoA	Precision	Recall
10%	300	262	90.80 ± 14.27	24.40 ± 2.87	0.94 ± 0.17	0.94 ± 0.16	0.98 ± 0.05
30%	400	388	72.40 ± 26.42	19.80 ± 2.99	0.98 ± 0.09	0.99 ± 0.08	0.99 ± 0.04
50%	500	446	51.20 ± 4.07	14.20 ± 2.86	0.97 ± 0.11	0.98 ± 0.09	0.98 ± 0.08

- `repeats_long`: 7 action blocks (one per gesture), each containing 8 repetitions of the same gesture;
- `repeats_short`: 7 action blocks (one per gesture), each containing 6 repetitions of the same gesture;
- `sequential`: 6 action blocks, each containing all the 7 gestures in sequence.

In particular, we focused on subject 03, a 37-year-old male, and on gestures n. 2 and 3, namely hand flexion and extension. As per [15,17–19], we pre-processed all signals beforehand using a 20 Hz to 700 Hz band-pass filter and several notch filters at 30 Hz, 50 Hz, 60 Hz, 90 Hz and 150 Hz to attenuate power-line noise.

Three decomposition approaches were tested:

Firing Rate Ordering. Two distinct decomposition models were employed, one calibrated on a signal recorded during the first session and the other on a signal recorded during the second session. In both cases the extracted MUs were ordered by their firing rate, from the highest to the lowest (the order was kept unchanged when performing inference).

Negative Entropy Ordering. As before, two distinct decomposition models were employed for the two sessions, but instead of being ordered by firing rate the extracted MUs were ordered by negative entropy (or neg-entropy), from the highest to the lowest (the order was kept unchanged during inference); the rationale behind neg-entropy ordering is that FastICA uses neg-entropy as contrast function to estimate MUAPTs, and thus the same MU in different recording sessions may have a similar neg-entropy value.

Re-calibration. A single decomposition model was employed: it was calibrated on a signal from the first session, and the estimated decomposition parameters (such as the mean vector, the whitening matrix, the separation vectors and the spike thresholds) were updated by re-calibrating the model on a signal from the second session, using a momentum $\beta = 0.5$.

In all the above tests, the same calibration signal was used, namely the slice from 50 s to 150 s of the `repeats_long` signal: such slice contains the sEMG recording during the performance of both flexion and extension gestures in sequence. We also tested multiple values for the hyper-parameter M_t, and the effects of re-sampling the sEMG signal at a lower frequency by applying an eighth-order type I Chebyshev filter and decimating at 2560 Hz and 1280 Hz (i.e., a half and a quarter of the original sampling frequency). The extension factor was fixed to $f_e = 8$ regardless of the sampling rate, since, as per [15,17], it was observed that the decomposition performance is similar for $f_e \in \{8, \ldots, 31\}$). The relationship between the sampling frequency, the number of target MUs and the actual number of extracted MUs can be observed in Fig. 1.

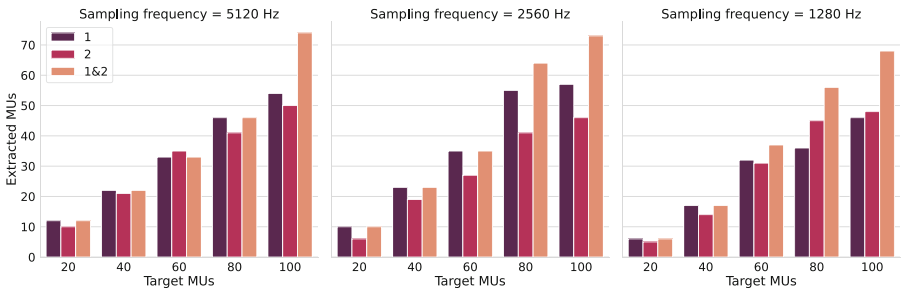


Fig. 1. Number of extracted MUs with respect to the number of target MUs obtained by calibrating the model on the first, second and both sessions (re-calibration).

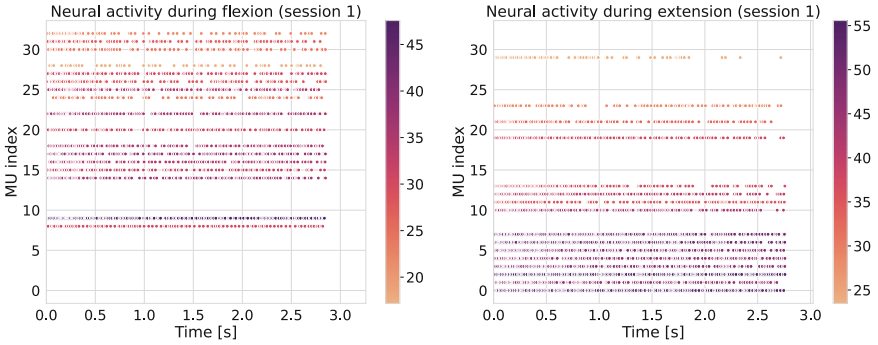


Fig. 2. MUs extracted by the calibrated decomposition model, color coded by firing rate. The firings characterizing a flexion gesture are shown on the left, those characterizing an extension are shown on the right: as it can be seen, the firing patterns are very different (in this example $M_t = 60$ and $f_s = 5120\,\text{Hz}$, but this difference in firing patterns was also observed for other values of M_t and f_s).

After calibration, the dataset for gesture classification was prepared as follows: (1) the slices corresponding to the gestures of interest from all the three sEMG signals (i.e., `repeats_long`, `repeats_short` and `sequential`) were considered, (2) since the duration of the gestures may differ, each slice was divided in several windows with a fixed length of 500 ms and with an overlap of 400 ms, (3) MUAPTs were extracted from each window using the calibrated decomposition model, (4) extracted MUAPTs were encoded as a $N \times S$ matrix of zeros and ones (the latter representing a spike), where N is the number of extracted MUs and $S = 0.5\text{s} \cdot f_s[\text{Hz}]$ is the number of samples, (5) the flattened MUAPTs were fed into a multilayer perceptron (MLP) binary classifier with a single hidden layer consisting of 32 neurons. We used a shallow classifier in order to avoid any data-driven, black-box feature extraction in addition to our BSS. In particular, we considered an (1) intra-session classification task, in which the MLP was trained on the flattened spikes extracted from the `repeats_long` and `repeats_short` signals recorded in the first session and was tested on the flattened spikes extracted from the `sequential` signal recorded in the first session, and (2) an inter-session classification task, in which the MLP was trained on the flattened spikes extracted from all the signals recorded in the first session and was tested on the flattened spikes extracted from all the signals recorded in the second session.

Regardless of the decomposition approach, the sampling frequency and the number of target MUs, the test accuracy in the intra-session task was 100%: this is reasonable, since flexion and extension involve different MUs and the decomposition model already separates them very neatly, as it can be observed in Fig. 2.

Concerning the inter-session task, performance varied depending on the approach used, on the sampling frequency and on the number of target MUs, as

shown in Fig. 3. In general, neg-entropy ordering seems to be more reliable than firing rate ordering, especially when there are only few extracted MUs (i.e., $M_t \in 20$) and the sampling frequency is high (i.e., $f_s \in \{5120, 2560\}$): this can be confirmed also by a visual inspection of the firing pattern obtained using both ordering (Fig. 4). As expected, the approach yielding the best accuracy overall is the re-calibration: since it extends the original separation matrix with new information from the new session, the ordering is guaranteed to be consistent. However, the position of the electrodes must be roughly the same between different recording session, otherwise the old separation vectors would become meaningless and may actually harm the decomposition.

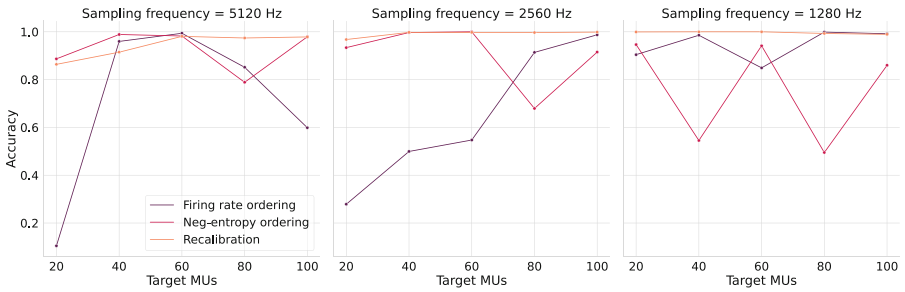


Fig. 3. Inter-session test accuracy by varying decomposition approach and number of target MUs.

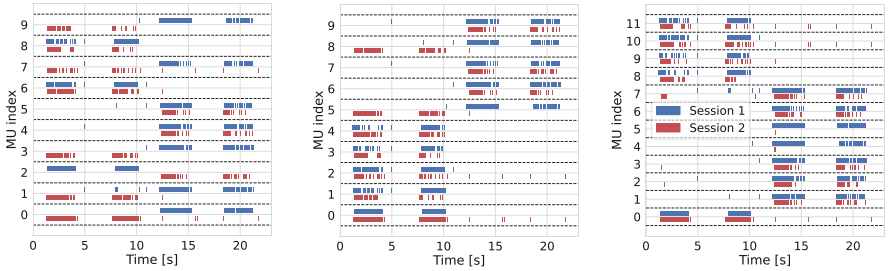


Fig. 4. Comparison between firing rate ordering (left), negative entropy ordering (middle), and re-calibration (right) with $M_t = 20$: the spikes were obtained by decomposing the same signals (i.e., first two flexion repetitions and first two extension repetitions from subject 03's `repeats_long` recording of both sessions).

4 Conclusion

In conclusion, we proposed three approaches to address the variability in the ordering of the MUs extracted using BSS-based methods when considering different recording sessions: two ordering criteria based on firing rate and negative entropy, and a re-calibration procedure, which allows the decomposition model

to retain information about previous recording sessions when decomposing new data. To demonstrate such approaches, we employed a pipeline based on convolutive ICA to decompose a sEMG signal into MUAPTs, which are then processed using a 1-layer MLP classifier to predict hand gestures. In particular, we showed that re-calibration is the most robust approach, as it yields an accuracy up to 99.4% in a 1-degree of freedom (DoF) classification setting. These results proved that our system, based on biological features rather than data-driven, black-box models, is capable of robust generalization.

Acknowledgements. The work was partially funded by the EU project BONSAPPS (g.a. 101015848).

References

1. Cannan, J., et al.: Human-machine interaction (HMI): a survey. In: University of Essex (2011)
2. Guo, L., et al.: Human-machine interaction sensing technology based on hand gesture recognition: a Review. IEEE Trans. Hum.-Mach. Syst. **51**(4), 300–309 (2021)
3. Wei, L., et al.: Gesture recognition using surface electromyography and deep learning for prostheses hand: state-of-the-art, challenges, and future. Front. Neurosci. **15**, 621885 (2021)
4. Zanghieri, M., et al.: sEMG-based regression of hand kinematics with temporal convolutional networks on a low-power edge microcontroller. In: 2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS), pp. 1–6 (2021)
5. Merletti, R., et al.: Tutorial. surface EMG detection in space and time: best practices. J. Electromyogr. Kinesiol. **49**, 102363 (2019)
6. Spiewak, C., et al.: A comprehensive study on EMG feature extraction and classifiers. Open Access J. Biomed. Eng. Appl. **1** (2018)
7. Chen, W., et al.: Hand gesture recognition using compact CNN Via surface electromyography signals. Sensors **20**(3), 672 (2020)
8. Zanghieri, M., et al.: Robust real-time embedded EMG recognition framework using temporal convolutional networks on a multicore IoT processor. In: IEEE Trans. Biomed. Circ. Syst. **14**(2), 244–256 (2020)
9. Masri, G., et al.: Design and control of a myoelectric prosthetic hand using multichannel blind source separation techniques'. In: 2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME), pp. 54–58 (2021)
10. Hyvarinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Netw. **10**(3), 626–634 (1999)
11. Holobar, A., et al.: Multichannel blind source separation using convolution kernel compensation. In: IEEE Trans. Sign. Process. **55**(9), 4487–4496 (2007)
12. Zanghieri, M., et al.: Temporal variability analysis in sEMG hand grasp recognition using temporal convolutional networks'. In: 2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), pp. 228–232 (2020)
13. Burrello, A., et al.: Tackling time-variability in sEMG-based gesture recognition with on-device incremental learning and temporal convolutional networks. In: 2021 IEEE Sensors Applications Symposium (SAS), pp. 1–6 (2021)

14. Gonzalo, G.A., Maekawa, K., Akazawa, K.: Decomposition of synthetic multi-channel surface-electromyogram using independent component analysis. In: Puntonet, C.G., Prieto, A. (eds.) Independent Component Analysis and Blind Signal Separation. ICA 2004. Lecture Notes in Computer Science, vol. 3195, pp. 985–992 . Springer, Berlin, Heidelberg (2004)
15. Negro, F., et al.: Multi-channel intramuscular and surface EMG decomposition by convolutive blind source separation. J. Neural Eng. **13**(2), 026027 (2016)
16. Thomas, J., et al.: Time-domain fast fixed-point algorithms for convolutive ICA. IEEE Sign. Process. Lett. **13**(4), 228–231 (2006)
17. Barsakcioglu, D., et al.: Control of spinal motoneurons by feedback From a non-invasive real-time interface. IEEE Trans. Biomed. Eng. 1–1 (2020)
18. Formento, E., et al.: Skilled independent control of individual motor units via a non-invasive neuromuscular-machine interface. J. Neural Eng. **18**(6), 066019 (2021)
19. Dai, C., et al.: Independent component analysis based algorithms for high-density electromyogram decomposition: systematic evaluation through simulation. Comput. Biol. Med. **109**, 171–181 (2019)
20. Jiang, X., et al.: Open access dataset, toolbox and benchmark processing results of high-density surface electromyogram recordings. IEEE Trans. Neural Syst. Rehab. Eng. **29**, 1035–1046 (2021)
21. Fuglevand, A.J., et al.: Models of recruitment and rate coding organization in motor-unit pools. J. Neurophys. **70**(6), 2470–2488 (1993)
22. Mohebian, M.R., et al.: Non-invasive decoding of the motoneurons: a guided source separation method based on convolution kernel compensation with clustered initial points. Front. Comput. Neurosci. **13**, 14 (2019)
23. Kaczmarek, P., et al.: putEMG - a surface electromyography hand gesture recognition dataset. Sensors **19**(16) (2019)

The Effect of Cognitive Load on Electrotactile Communication via a Multi-pad Electrode

Fabricio Ariel Jure[1], Erika Geraldina Spaich[1], Laura Petrini[2], and Strahinja Dosen[1](✉)

[1] Neurorehabilitation Systems, Department of Health Science and Technology, Faculty of Medicine, Aalborg University, Aalborg, Denmark
sdosen@hst.aau.dk

[2] Center for Neuroplasticity and Pain (CNAP), Department of Health Science and Technology, Faculty of Medicine, Aalborg University, Aalborg, Denmark

Abstract. Enhancing the perception of the surrounding environment by conveying information through tactile stimulation can improve performance, especially when other senses are already overloaded (e.g., first responders). Delivering electrotactile messages using spatiotemporal patterns has been suggested as an efficient and reliable approach to transmitting information through the skin. However, the effectiveness of this approach depends on the user's capability to recognize tactile messages in potentially challenging scenarios (e.g., rescue operations). Therefore, the present study evaluated the participants' performance in identifying electrotactile messages while they performed a parallel cognitive task. Thirty-six electrotactile messages were delivered through a 3×2 pad-matrix at the lateral torso in the baseline and multitasking conditions. The task-switching paradigm was employed as the parallel task, introducing cognitive interference to the electrotactile message recognition. The assumption was that the simultaneous task would impair the identification performance, but that this impairment could be minimized by decreasing the speed at which the message was delivered. The results indeed showed that the identification success rates decreased from ~83% in the baseline to ~61% during multitasking. Unexpectedly, the performance did not recover when the message duration was increased. Hence, modulating the message duration seems not to be an effective strategy to prevent lower performance while multitasking.

Keywords: Tactile stimulation · Sensory feedback · Cognitive flexibility · Multitasking

1 Introduction

In recent years, tactile stimulation technology has been gaining increasing attention in several application domains ranging from daily activities (e.g., mobile devices, sport tracking gadgets and video games) to biomedical applications (e.g., assisting sensory deprived users, feedback restoration in robotic prostheses and rehabilitation systems). Tactile sense can be used to convey information intuitively and without imposing an additional load on vision and audition [1, 2], and therefore, tactile feedback can enhance

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 530–539, 2023.
https://doi.org/10.1007/978-3-031-16281-7_50

the perception of the surrounding environment of the wearer leading to improved performance and reduced workload [3, 4]. There are different methods of conveying tactile information (e.g., employment of piezoelectric elements, pneumatic systems, hydraulic elements and vibration motors) [5–9]. Electrotactile technology can be used to develop compact solutions that can elicit rich tactile sensations, as the stimulation parameters (i.e., frequency, intensity, and location) can be finely and independently modulated [10, 11]. Electrical stimulation can be delivered through matrix electrodes with multiple pads turning this method into an attractive solution to achieve high-bandwidth tactile communication since the electrodes can be produced in different configurations (number, positioning and shape of the stimulation points), along with the possibility of modulating the different stimulation parameters of each pad.

The present work has been conducted within the framework of the SIXTHSENSE project, which focuses on the development of a biofeedback system to improve the efficacy and safety of first responders (e.g., firefighters, rescuers, paramedics, etc.) in hazardous situations where other senses are overloaded or partially deprived by the environment (e.g., vision impaired by smoke) [12]. The SIXTHSENSE project seeks to convey electrotactile feedback via a multi-pad electrode placed on the lateral side of the torso [10–12]. In a recent work [11], the authors have shown that the participants can reliably identify electrotactile messages encoded as spatial patterns delivered using sequential activation of pads (i.e., spatiotemporal approach). However, in that work, the participants were fully focused on message recognition, while in the envisioned application, the first responders will have to receive and interpret the messages while they attend to other activities, i.e., under physical and/or cognitive stress. The combination of multiple cognitive and/or motor tasks causes an impairment in the performance of related activities [13, 14]. Therefore, this study evaluated the identification of electrotactile messages delivered to the lateral torso of the body through a 3 × 2-pad matrix while participants were exposed to a simultaneous cognitively demanding activity. The assumption was that the message identification would be impaired while doing a simultaneous task, but that this impairment could be minimized by modulating the parameters of the feedback-encoding scheme. More specifically, we increased the stimulation duration per pad, effectively decreasing the speed at which the messages were delivered. Surprisingly, the results demonstrated that this approach was not effective and that other strategies need to be explored to achieve message transmission that is robust to the cognitive load.

2 Methods

Six healthy naïve participants took part in the study (2F-4M, age range 23–34 years). Informed, written consent was obtained from all participants before the experimental session. The protocol was approved by the ethical committee of Northern Jutland, Denmark (VN-20210055), and the experiment was performed in accordance with the Declaration of Helsinki.

2.1 Experimental Setup and Stimulation Intensity

The electrotactile stimulation was delivered using a multichannel stimulator (MaxSens, Tecnalia, Spain) controlled wirelessly by a host computer. A custom-made surface electrode (SIXTHSENSE ALPHA electrode, Tecnalia Serbia, Serbia [10])) was placed on the right, lateral side of the torso along the line connecting the armpit and the iliac crest at the hip [10, 11]. The electrode consisted of 8 cathodes (i.e., 6 circle-shaped and 2 rectangular with rounded corners), each of them surrounded by an anode forming 8 concentric cathode-anode pairs (Fig. 1A). Before positioning the electrode, the skin area was cleaned with alcohol swabs.

The electrotactile stimulation was delivered through the 3 × 2 pad matrix (i.e., 6 active pads), while the large pads (top and bottom) were not used. The multichannel stimulator generated current-controlled, biphasic and symmetric pulses with a pulse width set to 400 μs. The sensation thresholds (ST) were determined for each active pad using the method of limits [15]. The stimulation frequency was set to 10 Hz (i.e., the lowest frequency employed in the present study), while the amplitude was increased from 500 μA in steps of 100 μA until the participant indicated that he/she felt the stimulation. According to a previous study, the final stimulation intensity was set to 2 × ST [11]. Finally, the intensities were fine-tuned by increasing/decreasing the intensity in 100 μA-steps to evoke a similar sensation across the pads to avoid differences in saliency.

2.2 Encoding Method

To convey electrotactile messages, the spatiotemporal patterns were delivered by sequential activation of all the pads of the matrix. Each message was formed by ON pads (i.e., pads that form the pattern) and OFF pads (i.e., 'silent' pads). The ON pads were activated at one of three different frequencies, resembling different importance levels (i.e., low at 25 Hz, medium at 50 Hz or high at 150 Hz), whereas the OFF pads were always set to 10 Hz. In total, 36 different messages were designed (i.e., 12 spatial patterns × 3 frequency levels).

To cope with the possible cognitive overload during the multitasking paradigm, four different message durations were generated by varying the duration of the stimulation burst per pad, i.e., 750 ms, 1000 ms, 1500 ms, and 2000 ms. This corresponded to the total duration of 4.5 s, 6 s, 9 s and 12 s per message, respectively. As explained in the introduction, the assumption was that slowing down the speed at which a message was delivered would allow the participants to better receive and interpret the messages during the dual-tasking paradigm.

The sequential activation of the pads was always delivered following the same top-down order, i.e., from the top-back (pad 1) to the bottom-front (pad 6) of the pad matrix (Fig. 1A).

2.3 Experimental Procedure

A verbal explanation of the experimental procedure was given to the participants. The electrode was then positioned and the participants were asked to sit comfortably on a

chair in front of a screen. Afterward, ST was determined and the stimulation intensity was set for each pad. Four stimulation frequencies were presented to the participants (10, 25, 50, and 150 Hz) by randomly activating single pads until they were confident in recognizing the frequencies. The experiment consisted of a single session divided into training and validation stages, as explained below.

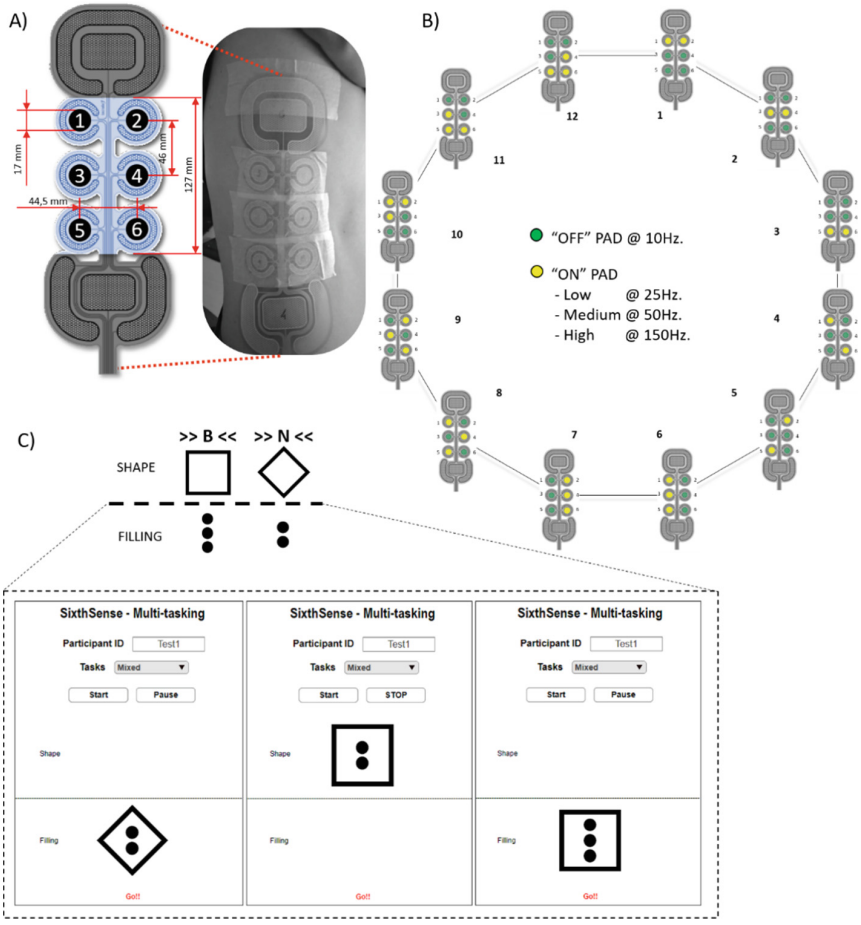


Fig. 1. Experimental setup. A) Electrode positioning at the lateral torso, the 3 × 2 pads matrix (highlighted) was used in this study. B) Electrotactile messages: spatial patterns and levels. The schematics of 12 spatiotemporal patterns. Each pattern could be delivered at three importance levels (high, medium and low) that were codified as different frequencies of the "ON" pads. C) Cognitive task based on the task-switching paradigm [16], depicting the rules and the custom-made GUI.

Training

The aim of this stage was to familiarize the participants with electrotactile messages. The familiarisation was organized in two phases, namely, learning and reinforcement.

Learning. The messages were sequentially delivered to the participant. All 12 patterns were first delivered at the frequency indicating low level, and then the procedure was repeated for medium and high levels. The participant received visual feedback indicating which pattern and level were transmitted. The graphical depiction of the patterns was shown to the participant during the learning process (Fig. 1.B).

Reinforcement. The messages were delivered randomly to the participant, and the participant had to verbally indicate the pattern (i.e., pronouncing the pattern number 1–12) and the level of importance (i.e., pronouncing "low", "medium" or "high") while a custom-made GUI provided visual feedback on the correct answer. The test was divided into three blocks of 12 messages, and the participants repeated the test until they could recognize the messages with an accuracy higher than 80% (i.e., at least 29/36 correct answers). The graphical depiction of the patterns was shown to the participants during this phase. During training, the stimulation duration was set at 1000 ms per pad.

Validation

In this stage, the aim was to determine the baseline success rate in recognizing the electrotactile messages. After this, the performance of message recognition was measured while the participants were exposed to a parallel cognitive task.

Message Identification - Baseline. The messages were delivered randomly to the participant. The test was divided into three blocks of 12 messages. The participants had 5 s to verbally indicate the pattern (i.e., pronouncing the pattern number 1–12) and the level of importance (i.e., pronouncing, "low", "medium" or "high"). The graphical depiction of the patterns was shown immediately after the end of stimulation, during the 5-s decision time, by a custom-made GUI. If the participant did not answer within those 5 s, the answer was considered wrong. A pause of 5–8 s was introduced before the next message was delivered, and a longer 1–2 min break was made between the blocks. This procedure was repeated for the four message durations randomly across participants.

Cognitive Task - Baseline. After participants finished with the message identification in the baseline condition, they were asked to perform a cognitive task. This task was based on a task-switching paradigm that tests the ability to process different types of information simultaneously [17, 18]. The test consisted of categorizing simple drawings according to a shape (i.e., a rhombus or a square) and according to the filling, which is the number of dots inside the shape (i.e., two or three dots) (Fig. 1C). Participants were instructed to react to the visual stimuli delivered by a custom-made GUI, by pressing a letter on the keyboard. For example, if the drawing appeared on the Shape side, participants had to answer with "B" for square or "N" for rhombus. If the drawing appeared on the Filling side, participants had to answer with "B" for three dots or "N" for two dots (see [16] for a review). The participants first trained the task by answering by Shape and then by Filling individually. After they felt comfortable with the task, 200 stimuli were randomly presented. The participants had 2 s to answer to each visual stimulus. If the participant did not answer within those 2 s, the answer was considered wrong.

Multitasking – Message Identification During Cognitive Task. The same procedure was followed as in the baseline assessment phase. Participants were instructed to pay attention to both tasks concomitantly, namely, the message identification and the task-switching test, following the same rules for each task as explained above (e.g., the graphical depiction of all patterns was presented to the participants after the delivery of electrotactile stimulation). After each block, habituation or sensitization to the stimulation was checked by asking the participants to assess the perceived stimulation sensation.

2.4 Data Analysis

The main outcome measures were the success rate (SR) in identifying the complete message and the percentage of task error (TE) in the task-switching test. The identification of the complete message was deemed successful if the participant correctly recognized the pattern and the importance level simultaneously.

The statistical tests were performed using IBM SPSS Statistics 27.0 (SPSS Inc., USA). A p-value < 0.05 was established as a threshold for statistical significance. The data were analysed using generalized linear mixed models (GLMMs). All models controlled for the within-participants variation by considering a random intercept and random slopes. Iterations were set at 100 and a criterion for convergence of $1{\cdot}10^{-6}$ was considered. A scaled identity covariance structure was used, in which variances are constant and no correlation was assumed between the elements. The sequential Sidak test was used for the correction of multiple comparisons.

The SR was analysed using a GLMM, with *Task* (*baseline* and *multitasking*), *Duration* (*750* ms, *1000* ms, *1500* ms, and *2000*) ms and their interaction as fixed factors, including the intercept. The model used a normal distribution and an identity link function (normality assessed with the Shapiro-Wilk test).

The TE was analysed using a GLMM, with *Condition* (*baseline, 750* ms, *1000* ms, *1500* ms, and *2000* ms) as a fixed factor, including the intercept. The model used a gamma distribution and a log link function (normality assessed with the Shapiro-Wilk test).

3 Results

The experimental session lasted approximately 2½ hours. All participants completed the experiment without reporting habituation nor sensitization to the stimulation. The maximum number of blocks in the reinforcement phase was 2 for all participants. The analysis indicated a strong effect of *Task* for the SR (GLMM: $F_{(1,34)} = 76.023$, $p < 0.001$). More specifically, when participants performed a concomitant cognitive task, the mean message identification SR dropped from 83.08% (1.94%) in the baseline to 60.83% (2.66%) during multitasking (pairwise $t = 8.719$, $p < 0.001$). However, the analysis did not show the effect of *Duration* (GLMM: $F_{(3,34)} = 1.390$, $p = 0.262$), nor any interactions (GLMM: $F_{(3,34)} = 0.091$, $p = 0.965$). Table 1 summarized the mean and the standard deviation for *Task* and *Duration.*

Figure 2 depicts the boxplots for the SR in the baseline (top) condition and during multitasking (bottom). The figure shows the SR when identifying the complete messages as well as marginal SRs when recognizing the pattern (level could be right or wrong) and level (pattern could be right or wrong).

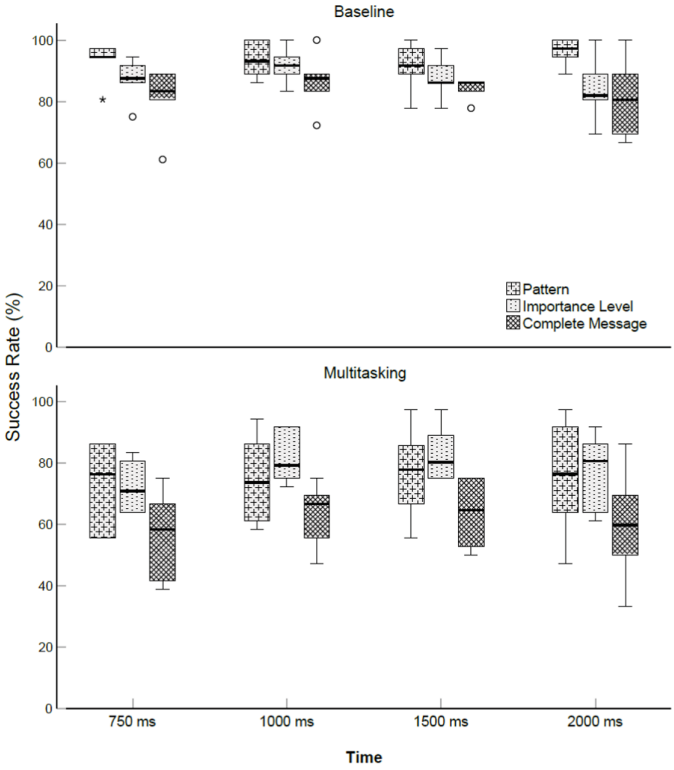


Fig. 2. Message identification success rates (SR). Boxplots represent the identification SR of the full message and the marginal SRs for recognizing the patterns and the importance levels individually. The top panel represents the baseline while the bottom panel is for multitasking condition. The x-axis represents the pad stimulation time (message duration).

Table 1. Descriptive statistics indicating the mean and standard deviation for the success rates in identifying the complete message, per *task* and *duration.*

Task	Duration			
	750 ms	1000 ms	1500 ms	2000 ms
Baseline SR (%)	81.02 (4.21)	86.57 (3.69)	83.88 (1.61)	81.02 (5.09)
Multitasking SR (%)	56.50 (5.80)	63.43 (4.14)	67.70 (4.50)	59.70 (7.27)

The statistical analysis for the cognitive task indicated a strong effect of *Condition* (GLMM: $F_{(4,25)} = 14.260$, $p < 0.001$). A post-hoc analysis specified that the task-switching performance was better (i.e., fewer mistakes) when participants were only dedicated to performing the cognitive task compared to multitasking, regardless of the message duration in the latter (*baseline* vs. all stimulation times: $p < 0.01$). No meaningful effects were found between the results for different message durations (Fig. 3).

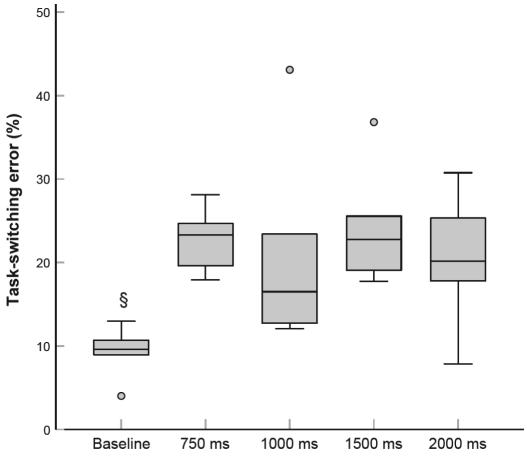


Fig. 3. The percentage of error during the cognitive task. Boxplots show the percentage of error when the participants performed the cognitive task for the baseline condition and while the participants were exposed to multitasking protocol (cognitive task simultaneous with the electrotactile message recognition). § indicates a significant difference ($p < 0.01$) between baseline and multitasking, regardless of the message duration.

4 Discussion

In this study, the identification of electrotactile messages was evaluated while participants were exposed to a simultaneous cognitively demanding activity. Thirty-six different spatiotemporal patterns were designed to convey 12 messages and three levels of importance. The electrotactile messages were encoded using a sequential - spatiotemporal - activation of all pads of the matrix. The messages were formed by ON pads (i.e., pads that carry the information) and OFF pads (i.e., 'silent' pads), and they were delivered to the participants in a baseline condition (i.e., without the simultaneous cognitive task) and in the multitasking paradigm (i.e., with a simultaneous cognitive task). The cognitive task was based on a task-switching approach, introducing cognitive interference while performing message recognition. Four different message durations (i.e., 4.5, 6, 9 and 12 s) were tested by modulating the stimulation time of each pad (i.e., 750, 1000, 1500 and 2000 ms per pad).

The results indicated that when the concomitant cognitive task was absent (baseline condition), the participants could identify the messages with high accuracy (mean SR ranging from ~81% to ~86%). The message duration did not have a notable effect on performance (Table 1). In line with a previous study [11], this observation highlights that electrotactile messages can be delivered reliably using the proposed spatiotemporal encoding scheme based on the sequential activation of active pads. Importantly, in the present study, the original encoding scheme was extended to include the level of importance encoded through the stimulation frequency. Therefore, the sequential encoding scheme can incorporate the information about both the pattern and the level of importance of the feedback variable communicated by the message, thereby increasing the communication bandwidth through the tactile channel. However, when a simultaneous cognitively demanding task was introduced, the performance of message recognition dropped substantially regardless of the message duration (mean SR ranging from ~56% to ~67% (Table 1)).

Even though the decrease in SR while performing a simultaneous task was expected, to our surprise, increasing the message duration did not improve the performance of message recognition during the multitasking paradigm. The hypothesis was that the participants would be better at recognizing the messages when the stimulation time per pad was increased, as they would have more time to focus on the pads and identify the state (ON or OFF). However, the cognitive task runs in parallel and continuously, and it could be that the cognitive load was such that the increased stimulation time was not effective assistance. Another reason could be that the participants successfully recognized the message, but the cognitive task interfered with their ability to memorize the pattern and/or the level of importance and recall it when they provided the answer.

The performance of the cognitive task was also substantially affected by the multitasking, regardless of electrotactile message duration. This could reflect the high cognitive demand, as the task required different mental operations, such as attention and working memory [19]. Consequently, switching, and multitasking operations might have been extremely cognitively costly in the present study to accomplish all the tasks successfully.

Therefore, the modulations of the stimulation time (message duration) seem not to be an effective strategy to improve message identification during cognitive overload, and thus, new strategies should be explored in future work. For instance, instead of prolonging the stimulation per pad, another strategy could be to fix the stimulation time but repeat the message more than once (“play” the same pattern repeatedly).

Acknowledgments. The current work was funded by the European Union’s Horizon 2020 research and innovation program under grant agreement No. 883315 (SixthSense).

References

1. Veen, V., Van Erp, J.B.F., Erp, V., et al.: Providing directional information with tactile torso displays. Proc. EuroHaptics 471–474 (2003)
2. Jones, L.A., Lockyer, B., Piateski, E.: Tactile display and vibrotactile pattern recognition on the torso. Adv Robot. **20**, 1359–1374 (2006)

3. Van Erp, J.B.F., Self, B.: Tactile displays for orientation, navigation and communication in air, sea and land environments (2008)
4. Van Erp, J.B.F., Werkhoven, P.: Validation of principles for tactile navigation displays. Proc. Hum. Factors Ergon. Soc. 1687–1691 (2006)
5. Phamduy, P., Rizzo, J.R., Hudson, T.E., et al.: Communicating through touch: macro fiber composites for tactile stimulation on the abdomen. IEEE Trans. Haptics **11**, 174–184 (2018)
6. Pohl, H., Brandes, P., Quang, H.N., Rohs, M.: Squeezeback: Pneumatic compression for notifications. In: Conference on Human Factors in Computing Systems - Proceedings, pp. 5318–5330. ACM, New York, NY, USA (2017)
7. Han, T., Anderson, F., Irani, P., Grossman, T.: HydroRing: supporting mixed reality haptics using liquid flow. In: Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, pp. 913–925. ACM, New York, NY, USA (2018)
8. Prewett, M.S., Elliott, L.R., Walvoord, A.G., Coovert, M.D.: A meta-analysis of vibrotactile and visual information displays for improving task performance. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **42**, 123–132 (2012)
9. Roady, T., Ferris, T.K.: An analysis of static, dynamic, and saltatory vibrotactile stimuli to inform the design of efficient haptic communication systems. Proc. Hum. Factors Ergon. Soc. **56**(1), 2075–2079 (2012)
10. Štrbac, M., et al.: Electrotactile stimulation, a new feedback channel for first responders. In: Ayaz, H., Asgher, U., Paletta, L. (eds.) Advances in Neuroergonomics and Cognitive Engineering. AHFE 2021. Lecture Notes in Networks and Systems, vol. 259. Springer, Cham, pp. 489–496 (2021)
11. Jure, F.A., Spaich, E.G., Malesevic, J., et al.: Encoding of spatial patterns using electrotactile stimulation via a multi-pad electrode placed on the torso. Artif. Organs (2022)
12. SIXTHSENSE: An EU H2020 research and innovation action. https://sixthsenseproject.eu/. Accessed 9 Feb 2021
13. Hyndman, D., Ashburn, A., Yardley, L., Stack, E.: Interference between balance, gait and cognitive task performance among people with stroke living in the community. Disabil Rehabil. **28**, 849–856 (2006)
14. Wang, X., Pi, Y., Chen, P., et al.: Cognitive motor interference for preventing falls in older adults: a systematic review and meta-analysis of randomised controlled trials. Age Ageing **44**, 205–212 (2015)
15. Kingdom, F.A.A., Prins, N.: Psychophysics: a practical introduction: second edition. Psychophys. A Pract. Introd. Second Ed. 1–331 (2016)
16. Stoet, G., O'Connor, D.B., Conner, M., Laws, K.R.: Are women better than men at multitasking? Econ Multitask, **1**(1), 19–32 (2016)
17. Li, B., Li, X., Stoet, G., Lages, M.: Exploring individual differences in task switching. Acta Psychol. (Amst) **193**, 80–95 (2019)
18. Hsieh, S., Gade, M., Kessler, Y., et al.: Efficient multitasking: parallel versus serial processing of multiple tasks. Front. Psychol. **6**, 1366 (2015)
19. Monsell, S.: Task switching. Trends Cogn. Sci **7**, 134–140 (2003)

TinyML: The Quest for Next-Generation Computing

Inter-Operability of Compression Techniques for Efficient Deployment of CNNs on Microcontrollers

Hamoud Younes[1,2(✉)], Hugo Le Blevec[1], Mathieu Léonardon[1], and Vincent Gripon[1]

[1] IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238 Brest, France
{hugo.leblevec,mathieu.leonardon,vincent.gripon}@imt-atlantique.fr
[2] Department of Computer and Communication Engineering, Lebanese International University, Bekaa, Lebanon
hamoud.younes@liu.edu.lb

Abstract. Machine Learning (ML) has become state of the art for various tasks, including classification of accelerometer data. In the world of Internet of Things (IoT), the available hardware with low-power consumption is often microcontrollers. However, one of the challenges for embedding machine learning on microcontrollers is that the available memory space is very limited, and this memory is also occupied by the rest of the software elements needed in the IoT device. The problem is then to design ML architectures that have a very low memory footprint, while maintaining a low error rate. In this paper, a methodology is proposed towards the deployment of efficient machine learning on microcontrollers. Then, such methodology is used to investigate the effect of using compression techniques mainly pruning, quantization, and coding on the memory budget. Indeed, we know that these techniques reduce the model size, but not how these techniques interoperate to reach the best accuracy to memory trade-off. A Convolutional Neural Network (CNN) and a Human Activity Recognition (HAR) application has been adopted for the validation of the study.

Keywords: CNN · Quantization · Pruning · Coding · Microcontrollers

1 Introduction

Convolutional neural Networks (CNNs) are state-of-the-art solution for computer vision tasks including image classification, object detection, etc. In the last decade we have been witnessing an increased adoption of CNNs in a wider range of applications such as voice recognition, near-sensor intelligence, human activity recognition and more. These applications demand the availability of smart devices that can perform inference in a faster and more energy efficient fashion [1]. However, CNNs are both computationally intensive and memory intensive, which makes it challenging for their deployment on the edge. This

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 543–552, 2023.
https://doi.org/10.1007/978-3-031-16281-7_51

is specifically the case for the deployment on memory-constrained devices such as Microcontroller Units (MCUs). The main challenge in embedding CNNs on microcontrollers is that the available volatile and non-volatile memories are in the order of few hundred/thousand kilobytes [2]. For applications on the edge, the available memory is also occupied by the rest of the software stack, e.g. WiFi, Bluetooth, etc.

One of the widely used methods to reduce the computational and memory requirements of CNNs is compression. Compressing a CNN can be achieved through a set of techniques such as quantization, pruning, and coding (in our work, coding refers to using lossless compression techniques when storing the network parameters into memory). In principle, quantization and pruning are known to reduce the memory footprint at the expense of accuracy reduction. As for coding, it reduces the CNN memory requirement at the expense of adding a latency overhead during inference to reconstruct the initial model parameters. Applying the aforementioned techniques is not a straightforward process as it is not known which quantization level to use, the amount of parameters to prune, how to combine two or more techniques, and if coding is necessary or not. Also, the existing deep learning frameworks fall short in some scenarios, for example it is not possible to use any number of bits for quantization.

In this paper, such challenges and questions are tackled through the following main contributions:

- A methodology towards embedding efficient CNNs on microcontrollers using several compression techniques and open source frameworks for deep learning training is proposed.
- A detailed study on the interoperability of compression techniques to achieve a high accuracy to memory trade-off is presented.

The rest of the paper is organized as follows: Sect. 2 describes the compression techniques used in the study. Section 3 details the proposed methodology for efficient CNN deployment on microcontrollers. Section 4 presents the experiment performed during the study including, the CNN architecture, training procedure, and the case study adopted for verification. Section 5 highlights the findings and shows a thorough comparison between the different simulated scenarios. Section 6 concludes the paper and illustrates on future work.

2 Compression Techniques

Quantization has shown consistent reliability and success in addressing the issues of deploying neural networks on constrained hardware platforms by reducing the number of bits required for the representation of each weight. Several works have tackled quantizing the training phase [3,4], while others tackled the quantization in the inference phase [5,6]. However, it has been shown that going below half-precision is a challenging task and requires a lot of fine tuning to maintain an acceptable error rate [7]. This challenge has provoked the need

for frameworks such as TensorFlow Lite Micro [8], Larq Compute Engine [9], CMSIS-NN [10], and CMIX-NN [11] for the use of 8-bit and below quantization for both training and deploying neural networks on microcontrollers.

Pruning is the process of systematically removing parameters/neurons from an existing network to reduce the overall size. Pruning methods usually differs in their choices regarding sparsity structure, scoring, scheduling, and fine tuning. Some methods adopt unstructured pruning i.e. prune individual parameters, others prune a fraction of the parameters with the lowest score (absolute value) within a layer or prune all desired weights at once, etc. [12]. During development, pruning can be applied using built-in functions in frameworks such as PyTorch and TensorFlow, or through a custom technique as the ones proposed in [13,14].

Lossless Compression (Coding) as its name implies, permits the reconstruction of the original data from compressed data without any loss in information. Lossless compression is widely used in "ZIP" file format and "gzip" tool in Linux environments. One of the widely used lossless coding methods is the "Huffman Coding" [15]. It uses a variable-length code-words to encode symbols based on the probability of occurrence of each symbol. Huffman coding has been applied to several state-of-art neural networks such as: ResNet-20, VGG-16, LeNet, etc. and results showed that a compression ratio up to 31× can be achieved for such architectures [16,17].

3 Methodology

Figure 1 presents the proposed methodology towards the efficient deployment of CNNs on microcontrollers with a set of compression techniques. This methodology will be followed in order to study the interoperability of different compression methods on the accuracy to memory trade-off. The process can be summarized in five phases:

- **P1**: Define a CNN model using PyTorch where the number of feature maps and kernel size can be altered at any time. This model will be the baseline used for float32 training. Then, a second model will be defined where each layer (convolution, linear, etc.) from the baseline model will be replaced with its quantized implementation. Existing frameworks (PyTorch, TensorFlow, etc.) offers quantization only for $nbits \in [8, 16]$. Thus, an ongoing open-source framework called *Brevitas* is adopted [18]. Brevitas allows the use of any number of quantization bits, offers the capability to quantize either weights or activations or both, and provides choosing any quantization method. However, using Brevitas for low bit quantization ($nbits = 1, 2$), the training process requires extensive fine-tuning to maintain an acceptable accuracy loss. Thus, a third model that extends Binary Connect [19] is defined. This model can quantize the weights to [−1,1] or [−1,0,1] for $1 - bit$ and $2 - bit$ quantization respectively.

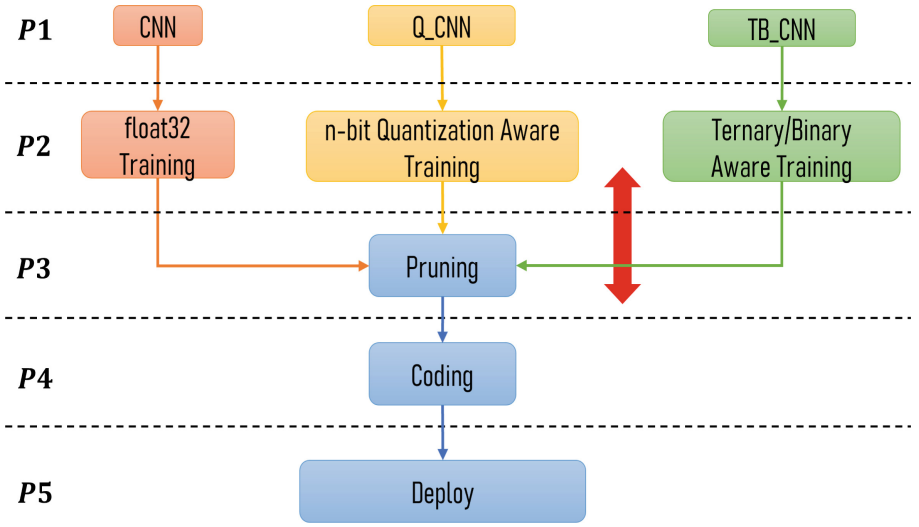


Fig. 1. Proposed methodology towards efficient CNN deployment on microcontrollers

- **P2**: Train all the models then fine tune them to obtain a comparable accuracy with respect to state-of-the-art solutions. As fine-tuning is not an easy task, techniques such as Neural Architecture Search (NAS) [20] can be used to define a set of hyperparameters among which some lead to the best possible accuracy.
- **P3**: Apply a pruning method on the pre-trained models and fine tune to achieve a similar accuracy compared to the baseline model. It is worth mentioning that for low-bit quantization, pruning a CNN based on the weights values tends to be less effective as the set of possible weight values is very small $(-1, 0, 1)$. A workaround could be to apply pruning before binarization.
- **P4**: For all the models, apply lossless compression using ZIP or gzip, which offers the same compression style as Huffman coding.
- **P5**: For deployment on microcontroller, a traditional deployment pipeline can be followed to obtain the flat buffer C array. Starting from a ".pth" file, Open Neural Network Exchange (ONNX) can be used to obtain a ".onnx" file, which is converted to a ".tflite" file. Finally, TFLite Micro library is used to export the C array. Note that for quantized models, Brevitas exports the weights in floating-point representation, thus it is necessary to apply post-training quantization to obtain the quantized weights.

Although frameworks such as PyTorch and Tensoflow offer out of the box pruning methods, they can only be combined with half-precision and 8-bit quantization. Moreover, CMIX-NN offers quantization with less than 8-bits, but the library is written in C and intended only for inference. Such problems are diminished with the use of the proposed methodology as it integrates several frameworks and provides insight on their capabilities. In addition, we applied this

methodology to a case study and the results are thoroughly explained and discussed in Sect. 5.

4 Experimental Setup

This section describes the experimental setup performed to obtain an understanding on the interoperability of compression methods on the memory requirements of CNNs. In this paper, a Human Activity Recognition (HAR) problem [21] is adopted for derivation of the study. Such problem has been one of the targets for embedded machine learning on microcontrollers among the tiny Machine learning (tinyML) community as it helps understanding the human movement behaviour and how it could be extended to robotics and industrial applications [22].

4.1 Case Study: Human Activity Recognition Using CNNs

In this work, we used the dataset hosted by the university of California (UCI-HAR) [23]. The dataset describes 6 daily life activities: walking, walking upstairs, walking downstairs, sitting, standing, and lying. Each activity is a one-dimensional time series of 2.56 s corresponding to 128 samples obtained 50 Hz sampling frequency. Each activity has 9 channels: 3 axes of total acceleration, 3 axes of angular velocity and 3 axes of body acceleration. The whole dataset contains a total of 10299 samples. For all training configurations, the UCI-HAR raw data has been normalized following the z-score of the training set.

As for the choice of CNN, a ResNet model with two blocks based on ResNet v1-6 presented in [24] is selected for the baseline model.

4.2 Training Procedure

The training process differs from one model to another due to the different representation of weights and hence different techniques could be required for fine tuning. However, there some notes to highlight:

- The kernel size has been fixed for all models to 3, while the number of feature maps (fmaps) is varied for 16, 32, 48, 64, and 96.
- L1-unstructured pruning from PyTorch has been applied to all models with a pruning rate $p \in [0.2, 0.4, 0.6, 0.8, 0.9, 0.95]$. This technique zeroes out the weights that have the lowest $l1$-norm.
- All models have been trained using 10-fold cross validation and the reported accuracy result is computed as the average among these runs.

Baseline Training. The model is trained for 300 epochs using a batch size of 64. SGD has been chosen as an optimizer with a learning rate of 0.05, momentum of 0.9, and a weight decay of 0.0005. A MultiStepLR scheduler is used to decrease the learning rate by a factor of 0.13 at epochs 100, 200, 250. Mixup [25] has been adopted for training with a factor of 0.2.

Quantized Training. The training procedure is similar to the baseline model with few changes. Training is performed through Quantization Aware Training (QAT) with uniform scale quantizer [26], where gradients are computed using the quantized versions of the weights, while non-quantized parameters are updated using them. For a weight value w, the quantized value w_q is defined as:

$$w_q = quantize(w, b, s) = clip(round(w_q * s), -2^{b-1} + 1, 2^{b-1} - 1) \quad (1)$$

where $nbits$ is the quantization bits, s is the quantization scale set as a learnable parameter, and the *clip* function is defined as:

$$clip(w, l, u) = \begin{cases} l & w < l \\ w & l \leq w \leq u \\ u & x > u \end{cases} \quad (2)$$

We trained three models where the quantization $nbits \in [4, 8, 16]$.

Ternary/Binary Training. For the training of these two models, the weight decay is reduced to 0.004 while preserving other hyperparameters from the baseline model. To compensate for the accuracy drop introduced from precision loss, batch normalization is applied after all convolution layers. Also, the model is trained for 120 epochs using float32 weights, then we clip the weights to [−1, 1]/[−1, 0, 1] for the remaining epochs. As for Mixup, it hasn't been used as the performance of low-bit quantized CNNs tends to be unstable.

5 Results and Discussion

Figure 2 presents the error rate (computed as $e = 1 - accuracy$, shown in log scale) versus the compressed model size for the models with different compression techniques. The compressed model size refers to the required non-volatile memory, which is to be uncompressed (decoded) during inference. It is worth mentioning that PyTorch framework keeps a value of zero in place of the pruned weights, thus the pruned model size is the same as the baseline. Hence, it is important to apply coding in order to get the real pruned model size. Moreover, models with a compressed model size greater than 400 KB (e.g. float32 and 16-bit models with 96 feature maps) have been left out as such size, when uncompressed, is already close to the maximum available volatile memory on mainstream microcontrollers. The plot shows the configurations with a compressed model size of 100 KB and less, such value is the target for most applications [1,27]. The following observations can be highlighted by analyzing the plot:

– In most of the configurations, quantization offers a reduced error rate compared to pruning for the same model size. As the number of feature maps decreases, the error rate increases rapidly when pruning is applied regardless of the quantization level.

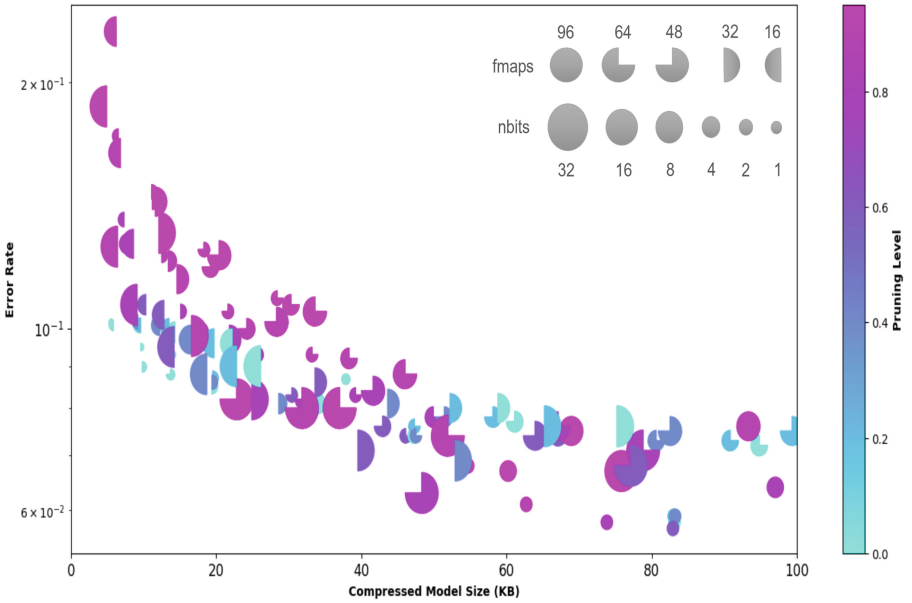


Fig. 2. Simulation results for models under different compression techniques

- Low-bit quantization could achieve similar model size compared to high pruning levels but with a much reduced error rate. For example, a binarized model has the same size as a 95% pruned float32 model for about half the error rate.
- The model size decreases as the pruning level increases for models with float, 16-bit, and 8-bit quantization. However, for 4-bit quantization, a high pruning level (>60%) must be applied to obtain a noticeable decrease in the model size. This is due to the fact that as the number of bits decreases, the possible weight values decrease, hence weight-magnitude-pruning techniques effect tends to be negligible.
- Reducing the number of feature maps often offers a better error to model size trade-off compared to increasing the pruning level. For instance, a model with 16 feature maps has a lower model size compared to a 95% pruned model with 96 feature maps, with a slight difference in the error rate.
- Models with small number of feature maps offer a smaller size compared to quantized models with $nbits > 2$. However, the error rate is lower when quantization is adopted.
- Going for a region with a model size less than 20 KB is accompanied with an high error rate. Typically, the models in this region have small number of feature maps, a high pruning level, and 4-bit or less quantization.
- Combining a quantization level n with a pruning rate p offers a smaller model size compared to a model with $n/2$ quantization level at the expense of increased error rate. Thus, as a trade-off, it is better to use a lower quantization level instead of a higher one with pruning, but with an increased number of feature maps.

- Regardless of the number of feature maps and quantization level, applying more than 60% pruning results in an huge increase in the error rate. Thus, the best pruning level to achieve an acceptable error rate is around 40%. The model size of such configuration is then obtained based on the selected quantization level.
- 16-bit and 8-bit quantized models offer a much reduced model size compared to float32 with a similar error rate. The same could be said about lower quantization levels but with a noticeable increase in the error rate.
- Combining pruning with low-bit quantization (1 and 2-bits), leads to huge increase in error rate, hence such configurations are not present in the plot. So, it is advisable to apply low-bit quantization after pruning.

Based on the target application and the available memory budget, one could find several models that fits the need. For instance, an 8-bit quantized model with 32 feature maps has the same size as a 16-bit quantized model with 60% pruning. However, the error rate is much lower for the 8-bit model.

6 Conclusion

This paper presented a methodology for quantization aware training accompanied with pruning and coding to reach an efficient model design for the deployment on microcontrollers. The methodology has been applied on a human activity recognition task to study the interoperability between different compression techniques. The study is thoroughly discussed to highlight various design strategies for an error to model size trade-off. The findings paves the way towards more extensive experimental setup in the future, including different task, CNN architecture, and quantizer types. Moreover, for real-time applications, the error to latency trade-off is another aspect that could be investigated.

Acknowledgement. This work has been partly supported by the grant ADEME PERFECTO 2021. The authors would like to thank the GoodFloow company for the financial support and technical guidance to this project.

References

1. Shafique, M., Theocharides, T., Reddy, V.J., Murmann, B.: TinyML: current progress, research challenges, and future roadmap. In: Proceedings - Design Automation Conference, vol. 2021-December, pp. 1303–1306. Institute of Electrical and Electronics Engineers Inc., December 2021. ISBN: 9781665432740
2. Sakr, F., Bellotti, F., Berta, R., Gloria, A.D., Doyle, J.: Memory-efficient CMSIS-NN with replacement strategy. In: Proceedings - 2021 International Conference on Future Internet of Things and Cloud, FiCloud 2021, pp. 299–303. Institute of Electrical and Electronics Engineers Inc., August 2021. ISBN: 9781665425742
3. Banner, R., Hubara, I., Hoffer, E., Soudry, D.: Scalable methods for 8-bit training of neural networks. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates Inc. (2018)

4. Choi, Y., Choi, J., El-Khamy, M., Lee, J.: Data-free network quantization with adversarial knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2020
5. Banner, R., Nahshan, Y., Soudry, D.: Post training 4-bit quantization of convolutional networks for rapid-deployment. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. D., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32. Curran Associates Inc. (2019)
6. Cai, Y., Yao, Z., Dong, Z., Gholami, A., Mahoney, M.W., Keutzer, K.: ZeroQ: a novel zero shot quantization framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020
7. Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M.W., Keutzer, K.: A survey of quantization methods for efficient neural network inference. arXiv Version Number: 3 (2021)
8. David, R., et al.: TensorFlow Lite Micro: Embedded Machine Learning on TinyML Systems, October 2020
9. Geiger, L., Team, P.: Larq: an open-source library for training binarized neural networks. J. Open Source Softw. **5**, 1746 (2020)
10. Lai, L., Suda, N., Chandra, V.: CMSIS-NN: efficient neural network kernels for Arm Cortex-M CPUs. ArXiv arXiv:1801.06601 (2018)
11. Capotondi, A., Rusci, M., Fariselli, M., Benini, L.: CMix-NN: mixed low-precision CNN library for memory-constrained edge devices. IEEE Trans. Circuits Syst. II Express Briefs **67**, 871–875 (2020)
12. Blalock, D., Gonzalez Ortiz, J.J., Frankle, J., Guttag, J.: What is the state of neural network pruning? In: Dhillon, I., Papailiopoulos, D., Sze, V. (eds.) Proceedings of Machine Learning and Systems, vol. 2, pp. 129–146 (2020)
13. Molchanov, D., Ashukha, A., Vetrov, D.: Variational dropout sparsifies deep neural networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, vol. 70 of Proceedings of Machine Learning Research, pp. 2498–2507. PMLR, August 2017
14. Tessier, H., Gripon, V., Léonardon, M., Arzel, M., Hannagan, T., Bertrand, D.: Rethinking weight decay for efficient neural network pruning. J. Imaging **8**, 64 (2022)
15. Huffman, D.: A method for the construction of minimum-redundancy codes. Proc. IRE **40**, 1098–1101 (1952)
16. Han, S., Mao, H., Dally, W.J.: Deep compression: compressing deep neural network with pruning, trained quantization and Huffman coding. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, Conference Track Proceedings, San Juan, Puerto Rico, 2–4 May 2016
17. Gajjala, R.R., Banchhor, S., Abdelmoniem, A.M., Dutta, A., Canini, M., Kalnis, P.: Huffman coding based encoding techniques for fast distributed deep learning. In: Proceedings of the 1st Workshop on Distributed Machine Learning, Barcelona, Spain, pp. 21–27. ACM, December 2020
18. Pappalardo, A.: Xilinx/brevitas (2021)
19. Courbariaux, M., Bengio, Y., David, J.-P.: BinaryConnect: training deep neural networks with binary weights during propagations. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS 2015, Montreal, Canada, Cambridge, MA, USA, pp. 3123–3131. MIT Press (2015)
20. Kyriakides, G., Margaritis, K.: An introduction to neural architecture search for convolutional networks. arXiv preprint arXiv:2005.11074 (2020)

21. Anguita, D., Ghio, A., Oneto, L., Parra Perez, X., Reyes Ortiz, J.L.: A public domain dataset for human activity recognition using smartphones. In: Proceedings of the 21th International European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, pp. 437–442 (2013)
22. Coelho, Y.L., Santos, F.A.S., Frizera-Neto, A., Bastos-Filho, T.F.: A lightweight framework for human activity recognition on wearable devices. IEEE Sens. J. **21**, 24471–24481 (2021)
23. Chetty, G., White, M., Akther, F.: Smart phone based data mining for human activity recognition. Procedia Comput. Sci. **46**, 1181–1187 (2015)
24. Novac, P.E., Hacene, G.B., Pegatoquet, A., Miramond, B., Gripon, V.: Quantization and deployment of deep neural networks on microcontrollers. Sensors **21**, 2984 (2021)
25. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. arXiv Version Number: 2 (2017)
26. Wu, H., Judd, P., Zhang, X., Isaev, M., Micikevicius, P.: Integer quantization for deep learning inference: principles and empirical evaluation. arXiv Version Number: 1 (2020)
27. Lin, J., Chen, W.-M., Cai, H., Gan, C., Han, S.: MCUNetV2: memory-efficient patch-based inference for tiny deep learning, October 2021

Energy Consumption Analysis of Pruned Semantic Segmentation Networks on an Embedded GPU

Hugo Tessier[1,2(✉)], Vincent Gripon[2], Mathieu Léonardon[2], Matthieu Arzel[2], David Bertrand[1], and Thomas Hannagan[1]

[1] Stellantis, Centre Technique Vélizy, 78140 Vélizy-Villacoublay, France
{hugo.tessier,david.bertrand,thomas.hannagan}@stellantis.com
[2] IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238 Brest, France
{vincent.gripon,mathieu.leonardon,matthieu.arzel}@imt-atlantique.fr

Abstract. Deep neural networks are the state of the art in many computer vision tasks. Their deployment in the context of autonomous vehicles is of particular interest, since their limitations in terms of energy consumption prohibit the use of very large networks, that typically reach the best performance. A common method to reduce the complexity of these architectures, without sacrificing accuracy, is to rely on pruning, in which the least important portions are eliminated. There is a large literature on the subject, but interestingly few works have measured the actual impact of pruning on energy. In this work, we are interested in measuring it in the specific context of semantic segmentation for autonomous driving, using the Cityscapes dataset. To this end, we analyze the impact of recently proposed structured pruning methods when trained architectures are deployed on a Jetson Xavier embedded GPU.

Keywords: Deep learning · Compression · Pruning · Hardware implementations · Energy

1 Introduction

In the course of the last decade, deep neural networks have become a staple in many computer vision tasks such as classification, detection or semantic segmentation. This makes them especially interesting for the field of autonomous vehicles. Cityscapes [3] is a semantic segmentation dataset designed for this field. It presents a significant challenge due to its high resolution, its size and the variety of images and classes.

For this dataset, state of the art is held by networks such as HRNets [20], which are composed of a large number of parameters and operations during inference and produce large intermediate products that cause a large memory footprint. Such costs are prohibitive for embedded devices and thus compromise their usage in autonomous vehicles.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 553–563, 2023.
https://doi.org/10.1007/978-3-031-16281-7_52

To reduce the cost of a network, the domain of neural networks compression presents multiple different methods, such as quantization [4], distillation [9] or pruning [5]. In this article we focus on pruning and more especially on "structured" filter pruning [13] that removes entire neurons in convolution layers of deep neural networks. Structured pruning can be a challenge to implement when applying different pruning rates for each layer of a network that includes long-range dependencies between layers because of, for example, residual connections [7]. Therefore, slimming down networks is a distinct issue, while most works in the field stick to putting weights to zero to measure, instead, the impact of pruning on a network's performance.

Our article aims at providing a solution to leverage irregular structured sparsity in HRNet-48 and measure the gains in energy consumption on a NVIDIA Jetson AGX Xavier embedded GPU. We report results using two different pruning methods of the literature and show how each of them allows for a different trade-off between performance and cost for the same network architecture. We also compare these results to that of non-pruned HRNet-32 and HRNet-18, and discuss the implications of these results in the general discussion on the ability of pruning to produce efficient architectures.

2 Related Works

2.1 Semantic Segmentation Networks

At first, deep semantic segmentation neural networks were built by concatenating a classification network and a decoder. The convolution layers of the (usually pretrained) encoder classification network produce intermediate representations that are successively fed to the decoder that upsamples and sums them together.

This principle was introduced by Long et al. [15] (FCN) and then perfected by Ronneberger et al. [18] (U-net) that cemented the symmetry between encoder and decoder that can be found in other networks, such as SegNet [1]. Dilated (or "atrous") convolutions [23] led to another family of semantic segmentation networks that, instead of upsampling and summing together intermediate representations of decreasing resolution, managed to alter a classification network so that said resolutions remains high. Such principle can be found among DeepLab networks [2], that also introduce the principle of pyramid pooling (i.e., operating on multiple resolutions at once to extract information at various scales).

Both ideas of pyramid pooling and maintaining high resolution gave birth to HRNet [20], which, instead of being a modified classification network, is fully designed for such a task. Variants of it, with an additional decoder, are state of the art [24]. However, HRNet tends to be a very complex network, whose cost is unsuitable for embedded hardware. Some works draw inspiration from HRNet to design networks more adapted to such hardware [10]. Our approach is to study how pruning could reduce the cost of HRNet.

2.2 Neural Network Pruning

First introduced during the late 80's [11], pruning is now a widespread method to reduce the cost of deep neural networks [6]. Since removing weights in an unstructured manner leads to sparse tensors that are difficult to optimize efficiently [12,16], it is possible instead to remove whole neurons, i.e. filters in the case of convolution layers [13]. Indeed, such kind of removal alters input and output dimensions of layers and it is necessary to remove inter-dependant weights between layers in a consistent way.

This is rather straightforward to do in simple networks such as VGG [19], but networks such as HRNets present much more intricate dependencies between layers, partly because of residual connections [7]. Indeed, such connections imply a certain regularity in the dimensions of intermediate products of the network. In the case of global pruning, where each layer can be pruned at different rates to get a global count of parameters (in opposition to local pruning that prunes equally every layer), pruning can break this regularity, which means that summing these tensors is not possible without scattering them into tensors of the same dimension. Multiple works solve this problem using different kind of operators [8].

3 Method

Our work focuses on pruning HRNet. Indeed, both its significance in the state of the art and its prohibitive cost in memory and computation power make it an especially relevant subject for experimentation to measure how pruning can make semantic segmentation suitable for embedded hardware.

We used the Cityscapes dataset [3] to train our networks. Indeed, Cityscapes is a popular dataset in the scope of autonomous vehicles and features high resolution images and annotations for 19 different classes (among others that are not used during training). All the results featured in this article concern networks we trained ourselves.

We chose two different structured pruning methods: Slimming [13], an influential method in the literature, and SWD [22], that is more recent and has shown a certain ability to give competitive results on multiple networks or tasks with different kinds of pruning structures.

All our measurements were run on a NVIDIA Jetson AGX Xavier embedded GPU under the "30W All" mode, with JetPack SDK 5.0, CUDA 11.4.14, cuDNN 8.3.2, TensorRT 8.4.0 EA and running inference with ONNX Runtime 1.12.0 while using the TensorRT execution provider and using the tegrastats utility to measure power consumption. The number of MAC operations is counted using the ONNX Operations Counter tool.

3.1 Training Conditions

Our HRNet networks are trained during 200 epochs with a batch size of 10 on 3 NVIDIA Quadro K6000 GPUs, using the Pytorch 1.9.0 framework with CUDA

11.3. We use the SGD optimizer with weight decay set to $5 \cdot 10^{-4}$, momentum set to 0.9 and base learning rate set to 0.01. We decrease learning rate using the poly policy with the power of 2 (i.e. learning rate is reduced by $(1 - \frac{current_epoch}{epochs})^2$ at each epoch). We use the RMI loss [25] and report accuracy in term of mean intersection over union (mIoU). Most of these hyperparameters come from the original papers of HRNet.

During testing, input images are of size $3 \times 1024 \times 2048$. During training, they are randomly cropped and resized, with a scale of $[0.5, 2]$, to $3 \times 512 \times 1024$. Data augmentation also involves random flips, random Gaussian blur, and color jittering. All samples from the dataset are normalized. More implementary precisions can be found consulting the source code available at https://github.com/HugoTessier-lab/Neural-Network-Shrinking.git.

Our baseline performance for HRNet-48 shows significant differences with the results announced in the original paper [20]. Indeed, using recent versions of Python and Pytorch (that we need for our tools to work properly), we were unable to reproduce the original results, even when applying the same training conditions or even when using available pretrained networks and the original source code. When using a pretrained HRNet-48 on Cityscapes, the best results we got were obtained with our implementation and lead to a mIoU accuracy of 73%. Therefore, we conclude that the performance we get with our own training, which is of 77% in mIoU, is good enough not to harm the conclusions of our article in any way.

3.2 Pruning Strategies

Pruning Structure and Target: We prune convolution filters and all the weights that depend on them. This includes corresponding batch normalization coefficients and kernels in filters of following layers. We also take into account long-range dependencies introduced by residual connections. The tensors of the network's weights are then reshaped so that the remaining parameters allocated in memory, the pruning target and the count of weights that still contribute to the network's function perfectly match.

Pruning Criterion: We prune filters on the basis of the magnitude of the learnt multiplicative coefficient in its corresponding batch-normalization layer [13].

Pruning Methods

- **Slimming**: During training, batch-normalization layers are penalized by a smooth-$\mathcal{L}_1$ norm of their multiplicative coefficient, in order to enforce sparsity. Networks are pruned in three steps, with a linearly increasing pruning rate until the final one is reached. After each pruning step, the network is fine-tuned during 20 epochs. Such an iterative process helps preventing layer collapse [21]. Once the final pruning rate is reached, the network is re-trained following a warm-restart schedule, which can be called LR-Rewinding [17].
- **SWD**: Since SWD prunes progressively throughout pruning, it does not need iterations. Instead we prune directly at the intended final pruning rate once

the networks have been trained under the constraint of SWD's penalty. The hyperparameters a_{min} and a_{max} of SWD are set to $1 \cdot 10^{-1}$ and $1 \cdot 10^{10}$ respectively, because with weight decay set to $5 \cdot 10^{-4}$, learning rate to 0.01 and the aforementioned scheduler, we get $a_{min} \times WD \times LR = 0.1 \times WD \times LR < WD \times LR$ and $a_{max} \times WD \times LR \times \prod_{n=1}^{200} \left(1 - \frac{n}{200}\right)^2 \approx 1 \cdot 10^{10} \times 1.25 \cdot 10^{-10} \approx 1$, which means that SWD starts off as negligible and ends up applying a penalty to targeted weights that almost equals their magnitude, thus pruning them. After that, LR-Rewinding is applied.

3.3 ONNX Implementation Details

The ONNX format has the clear advantage of being supported by a wide array of frameworks, that themselves can be ported on many different hardware. Therefore, gains observed in our experiments can be reproduced on other devices.

Structured pruning can introduce some dimensional discrepancies, which can be solved by introducing operators to reshape tensors coming from layers that were pruned differently [8]. In our implementation, we chose to insert a *ScatterND* ONNX operator, that inserts the data from a tensor into another whose dimension matches that of the layers that will take this tensor as an input. This operator is placed before any tensor addition operation to prevent any incompatibility in dimensions. This solves the problem but counts a drawback, that can be however put into perspective. Indeed, this operator needs additional temporary empty feature maps before summation, which introduces an additional cost in memory space and latency at inference. This problem is inevitable with our paradigm and can be found in previous works [8]. Preventing this problem would require rearranging the network in a much more complicated way, which we wished not to dwell into yet.

This implementation is the one that gave us the best performance in inference time, energy consumption and size of the network. However, it still has a non-negligible impact on performance, as can be seen in our experiments. This cost is likely due to the unconventional nature of this operator in neural networks, that makes it poorly supported by most frameworks. Indeed, TensorRT treats it as a "foreign node" that is accelerated using a library, Myelin, distinct from those used for all other nodes. The fact that the *Scatter* operator and its variants are only supported since TensorRT 8.0 hints that the acceleration of these operations still has room for improvement. Yet, this overall configuration is the one that performed the best during our preliminary experiments.

4 Experiments

We measured power consumption using the tegrastats utility every second while ONNX Runtime executes 1000 inference of a given network on GPU, after 100 inferences as a warm-up. We compute the integral of instantaneous power consumption (in mW) for all the duration of the experiment and get the overall

energy consumption (in J). We report the results for inputs of different resolutions and batch sizes in Fig. 1. We chose 2 different dimensions: $(1 \times 3 \times 64 \times 128)$ and $(1 \times 3 \times 512 \times 1024)$. We also report results for non-pruned baselines such as HRNet-32 or HRNet-18.

Figure 2 shows various trade-offs, for inputs of size $(1 \times 3 \times 512 \times 1024)$, between mIoU accuracy of the networks on CityScapes and their energy consumption, amount of remaining operations or parameters. All our reported numbers of operations do not take into account those of the *ScatterND* operators, but only that of convolutions. Indeed, as energy consumption already includes the impact of these operators, our operation count rather gives an indication on the distribution of sparsity in the network.

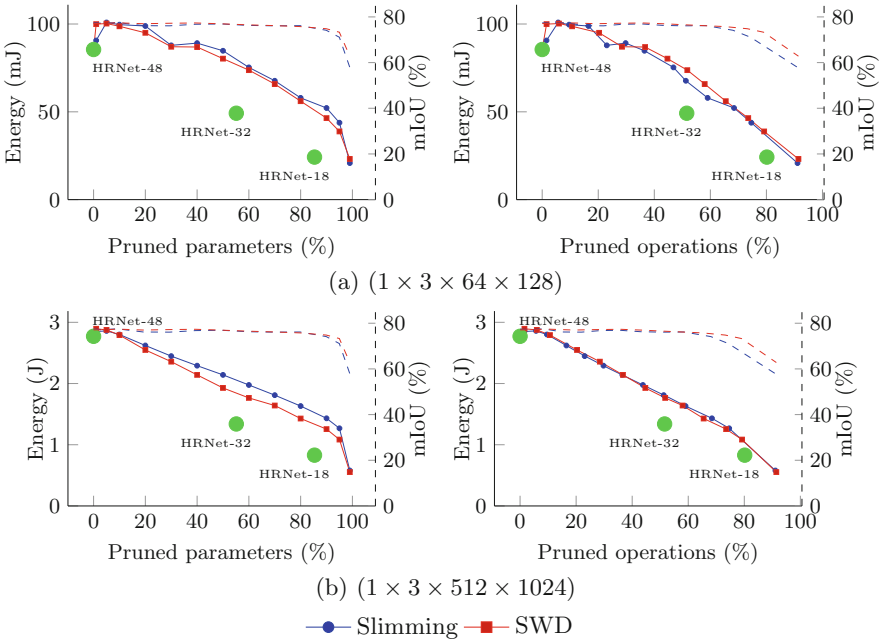


Fig. 1. Energy consumption, averaged over 1000 inferences, of HRNet-48, pruned at different rates. Dashed curves provide respective mIoU for each method. Green dots provide energy consumption of non-pruned network as a reference. Their position on the x-axis depends on their number of parameters (or operations) relatively to that of a non-pruned HRNet-48.

5 Discussion

5.1 Impact on Hardware

Our experiments show that we can leverage global filter pruning and reduce energy consumption. The efficiency of pruning slightly depends on the size of the inputs, even though results get more stable once inputs are of sufficient resolution, as can be seen in Fig. 1. This was achieved using *ScatterND* ONNX operations, whose cost in latency and energy is not negligible. Figure 1 shows that non-pruned smaller networks, such as HRNet-32 and HRNet-18, are energetically more efficient than pruned networks, even when comparing at equal number of operations. One would expect operations, latency and energetic consumption to be correlated, and that the difference between, for example, the consumption and latency of HRNet-32 and that of pruned networks of similar count of operations should match. However, this is only partly true, as indicated in Table 1.

Indeed, once the cost of the "foreign nodes", that are what TensorRT turn the *ScatterND* operations into, is removed from the overall latency of the network, results in latency are closer to each other but there is still a difference. When isolating the latency of convolution layers, that are supposed to count the same number of operations across the three compared networks, one can see that those of non-pruned HRNet-32 run significantly faster. As the difference in operations, that is minimal, or parameters, that should be in favor to the pruned networks, cannot explain this gap, we suppose that it is due to differences in the acceleration tactics empirically chosen by TensorRT. This would mean that not only *ScatterND* operations are costly but also that their presence harms the optimization of the rest of the network. This is likely due to their unconventional nature in convolution networks.

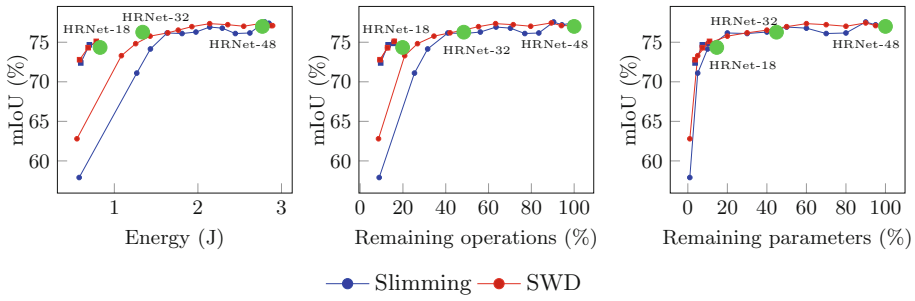


Fig. 2. mIoU accuracy of pruned HRNet-48 (filled circles) and HRNet-18 (squares) on CityScapes, for inputs of size ($1 \times 3 \times 512 \times 1024$), as a function of energy (left plot), remaining operations (middle plot) and remaining parameters (right plot). Green dots provide performance of non-pruned HRNet-48, HRNet-32 and HRNet-18 as a reference. Percentages of remaining operations or parameters are indexed on the cost of a non-pruned HRNet-48, whatever the network.

Table 1. Count of remaining parameters (Param.) and operations (Op.), overall latency (Lat.), latency with the exclusion of foreign nodes (Lat.*) and latency of convolution layers only (C. Lat.), for two different sizes of inputs, of HRNet-48 pruned using SWD (SWD) or Slimming (Slim.) and a non-pruned HRNet-32 (HRN32). For HRNet-32, the percentage of parameters and operations gives a relative comparison to those of non-pruned HRNet-48. Measurements and profiling were performed using the *trtexec* utility.

	Param.	Op.	Lat.	Lat.*	C. Lat.	Lat.	Lat.*	C. Lat.
			$(1 \times 3 \times 64 \times 128)$			$(1 \times 3 \times 512 \times 1024)$		
SWD	40.0%	48.2%	25.1 ms	20.3 ms	19.9 ms	245.2 ms	196.0 ms	188.0 ms
Slim.	30.0%	48.8%	22.5 ms	17.9 ms	17.5 ms	248.1 ms	199.8 ms	191.9 ms
HRN32	44.9%	48.4%	15.1 ms	15.1 ms	14.4 ms	158.3 ms	158.3 ms	147.7 ms

The most likely solutions to this problem would be either: 1) a better choice of operators, that we do not know as our solution is the one that gave us the best performance, 2) a new custom operator that cuts the drawbacks of our current solution, that uses preexisting ONNX operators or 3) a better support of *Scatter* operations and of this type of use in networks from TensorRT and other frameworks.

5.2 Pruning as Architecture Search

Figure 2 provides three different trade-off curves, for both pruned HRNet-48 and HRNet-18, which each lead to a different observation. Green dots provide performance of non-pruned networks. The accuracy-to-energy trade-off as well as the accuracy-to-operations trade-off is in favor of HRNet-18 compared to HRNet-48. On the contrary, HRNet-48 is better on the accuracy-to-parameters trade-off. In the literature on network pruning, the parameters count is often used to measure the complexity of the pruned network. Our results show that it is a bad proxy for its actual energy consumption and computational complexity. The reason is that, since different layers in the network are applied to intermediate representations of different resolutions, all parameters in the network do not have the same cost in term of operations and energy.

It seems that pruning HRNet-48 leads to save parameters that are responsible for many operations. Indeed, HRNet-18, that is uniformly thinner than HRNet-48, counts much fewer operations than pruned networks that contain as many parameters. This is surprising, since the criterion described in Sect. 3.2 does not explicitly imply any bias towards targeting parameters responsible for fewer or more operations. Since this behavior leads to architectures that seem sub-optimal, one can wonder if this the best that pruning can do or if it is due to some unexpected bias in the criterion.

One hypothesis would be that parameters that are responsible for many operations are indeed more important and of greater magnitude, in which case the count of operations and accuracy would be strongly correlated. This hypothesis

is rather unlikely. Indeed, one example of strong correlation is the parameters-to-accuracy trade-off in Fig. 2: all points are roughly located on the same average curve, whatever the pruning method or the pruned network. In the case of the operations-to-accuracy (or even the energy-to-accuracy) trade-off, there is a noticeable gap between pruned HRNet-48 and HRNet-18 as well as between networks pruned using SWD or Slimming. Moreover, networks pruned using SWD both contain fewer operations and have a better accuracy compared to those pruned using Slimming, even though they contain the same number of parameters, which is the opposite behavior to the one to be expected if the hypothesis was true.

If such parameters are not more important, then there is apparently no reason why pruning would prioritize saving them. Therefore it is likely that such a behavior is an accident and due to an unwanted bias in the criterion. Various types of analogous imbalances are to be seen in many criteria in the literature [21]. Since this bias visibly harms pruning in a way that makes it less efficient than training thin baselines from scratch, it seems that further investigations are necessary to figure out the reason of such biases and how to avoid them to improve considerably the performance of pruning.

Finally, the fact that pruning HRNet-18 leads to better results than pruning HRNet-48 leads to both the intuitive observation, that it is better to prune a network that is closer to the desired target, and the counter-intuitive observation that pruning the bigger network failed to come up with the same efficient architecture, while the ability of pruning to produce them is supposedly its main interest [14].

6 Conclusion

We have measured the energetic consumption of the HRNet-48 and HRNet-18 semantic segmentation networks, pruned using global structured pruning methods. Thanks to the use of *ScatterND* ONNX operations, we are able to leverage the sparsity on embedded GPU, demonstrating the ability of global pruning to reduce the cost of networks. However, our experiments also show the necessity both to elaborate a more efficient implementation of operators required to leverage pruning and to explore further how biased are pruning criteria, as they lead to distributions of sparsity that are energetically sub-optimal.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
2. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)

3. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
4. Courbariaux, M., Bengio, Y., David, J.-P.: BinaryConnect: training deep neural networks with binary weights during propagations. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
5. Han, S., Mao, H., Dally, W.J.: Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. arXiv preprint arXiv:1510.00149 (2015)
6. Han, S., Pool, J., Tran, J., Dally, W.: Learning both weights and connections for efficient neural network. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. He, Y., Zhang, X., Sun, J.: Channel pruning for accelerating very deep neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1389–1397 (2017)
9. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2(7) (2015)
10. Hong, Y., Pan, H., Sun, W., Jia, Y.: Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. arXiv preprint arXiv:2101.06085 (2021)
11. LeCun, Y., Denker, J., Solla, S.: Optimal brain damage. In: Advances in Neural Information Processing Systems, vol. 2 (1989)
12. Liu, Z.-G., Whatmough, P.N., Zhu, Y., Mattina, M.: S2TA: exploiting structured sparsity for energy-efficient mobile CNN acceleration. arXiv preprint arXiv:2107.07983 (2021)
13. Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., Zhang, C.: Learning efficient convolutional networks through network slimming. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2736–2744 (2017)
14. Liu, Z., Sun, M., Zhou, T., Huang, G., Darrell, T.: Rethinking the value of network pruning. arXiv preprint arXiv:1810.05270 (2018)
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
16. Ma, X., et al.: Non-structured dnn weight pruning–is it beneficial in any platform? IEEE Trans. Neural Netw. Learn. Syst. (2021)
17. Renda, A., Frankle, J., Carbin, M.: Comparing rewinding and fine-tuning in neural network pruning. arXiv preprint arXiv:2003.02389 (2020)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
20. Sun, K., et al.: High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514 (2019)
21. Tanaka, H., Kunin, D., Yamins, D.L., Ganguli, S.: Pruning neural networks without any data by iteratively conserving synaptic flow. In: Advances in Neural Information Processing Systems, vol. 33, pp. 6377–6389 (2020)

22. Tessier, H., Gripon, V., Léonardon, M., Arzel, M., Hannagan, T., Bertrand, D.: Rethinking weight decay for efficient neural network pruning. J. Imaging **8**(3), 64 (2022)
23. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
24. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12351, pp. 173–190. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58539-6_11
25. Zhao, S., Wang, Y., Yang, Z., Cai, D.: Region mutual information loss for semantic segmentation. In: Advances in Neural Information Processing Systems, vol. 32 (2019)

A Tiny CNN for Embedded Electronic Skin Systems

Fouad Sakr[1,2(✉)], Hamoud Younes[3,4], Joseph Doyle[2], Francesco Bellotti[1], Alessandro De Gloria[1], and Riccardo Berta[1]

[1] Department of Electrical, Electronic and Telecommunication Engineering and Naval Architecture, University of Genoa, Genoa, Italy
fouad.sakr@edu.unige.it,
{francesco.bellotti,alessandro.gloria,riccardo.berta}@unige.it
[2] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK
j.doyle@qmul.ac.uk
[3] IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238 Brest, France
hamoud.younes@imt-atlantique.fr
[4] Department of Computer and Communication Engineering, Lebanese International University, Bekaa, Lebanon
hamoud.younes@liu.edu.lb

Abstract. The quest for efficient Tiny Machine Learning on Microcontroller Units is increasing rapidly due to the vast application spectrum made possible with the advancement of Tiny ML. One application area that could benefit from such advancement is Electronic Skin systems, that are employed in several domains such as: wearable devices, robotics, prosthesis, etc. An e-skin system demands hard constraints including real-time processing, low energy consumption, and low memory footprint. This paper presents a tiny Convolution Neural Network (CNN) architecture suitable for the deployment on an off-the-shelf commercial microcontroller in compliance with the e-skin requirements. The training, optimization, and implementation of the proposed CNN are presented. The CNN implementation is optimized through layer fusion and buffer re-use strategies for efficient inference on edge devices. As a case study, experimental analysis of a touch modality classification task demonstrates that the proposed CNN-based system is capable of processing tactile data in real-time directly near the source while reducing the model size by up to 65% with respect to comparable existing solutions.

Keywords: Tiny ML · Convolution Neural Networks · Microcontrollers · Electronic skin · Tactile processing

1 Introduction

Tiny Machine Learning (Tiny ML) is an evolving paradigm that proposes the integration of Machine Learning (ML) algorithms with tiny hardware devices

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 564–573, 2023.
https://doi.org/10.1007/978-3-031-16281-7_53

Fig. 1. Electronic skin: blocks and functionality

mainly Microcontroller Units (MCUs). The aim of such integration is to utilize intelligence with an ultra-low power and low-cost efficiency for embedded systems on the edge. This is paving the way to the deployment of novel applications using battery-powered devices without the power-hungry processing support from the cloud [1]. To this extent, Tiny ML has been adopted in several application domains such as healthcare, smart cities, industrial control, etc. [2]. However, new application domains could join the rapid adoption of Tiny ML, specifically those with memory and power constraints such as the design of an efficient Electronic Skin (e-skin) system. These systems are widely used in wearable, prosthesis, and robotics applications.

An e-skin system, sketched in Fig. 1, consists of a set of distributed tactile sensors integrated with an embedded electronic system for tactile data decoding [3]. An array of tactile sensors detects and convert a mechanical stimuli into electrical signals. Then, the electronics interface applies signal conditioning and analog to digital conversion. Finally, an embedded processing unit (EPU) decodes the received tactile data using an intelligent mechanism. ML algorithms are the main candidate for the decoding process as they can extract meaningful information such as texture, pattern, etc. However, the computational complexity of ML algorithms, accompanied by the design requirements of an e-skin, poses a challenge for EPUs to support real-time operation with very low energy consumption. In this context, we propose a tiny Convolution Neural Network (CNN) model to be employed inside an EPU based on a Cortex-M MCU devoted to tactile data processing. The main contributions of this paper are thus:

- A Tiny Convolution Neural Network (TinyCNN) is proposed for processing tactile data. This model provides up to 10% higher classification accuracy compared to existing solutions with a reduced number of parameters and floating point operations (FLOPs).
- A platform-independent C implementation for the proposed TinyCNN, is provided for embedded processing. The implementation supports layer fusion and buffer reuse to increase memory and latency efficiency.
- The proposed CNN architecture is deployed on Cortex-M MCU and verified for touch modality classification. The obtained results lead to a real-time operation [4] with 0.5 ms and relatively low energy consumption of 65.36 µJ.

The rest of the paper is organized as follows: Sect. 2 provides an overview of ML methods used in the literature for tactile data processing. Section 3 describes the implementation and optimization of TinyCNN on Cortex-M microcontroller and defines the performance metrics used. Section 4 presents the experimental results of an TinyCNN implementation for e-skin touch modality classification, and compares them with existing solutions. Section 5 concludes the paper and proposes future work.

2 Machine Learning for Tactile Data Processing

2.1 Related Work

A variety of ML algorithms have been adopted in literature for tactile data processing in applications such as: surface texture, slip detection, object recognition, and touch modality classification. A tactile sensory array has been employed on a humanoid for object structure identification (e.g. soft, hard, etc.) [5]. The identification has been performed using a k-Nearest Neighbor (k-NN) classifier. Support Vector Machine (SVM) has been used to allow another humanoid to categorize ten objects (glass, sponge, etc.) using their surface texture in [6]. Another SVM classifier has been utilized in [7] that allows a robot to move objects from one place to another with a stable slip and grasp detection mechanism within 30 ms.

For the touch modality classification task presented in [8], different types of models have been employed, such as SVM based on tensorial kernel [8], k-NN [9], and a deep CNN [10,11]. Authors in [12] have proposed an SVM accelerator on Virtex-7 Field Programmable Gate Array (FPGA) that can classify an input touch modality of size $4 \times 4 \times 20$ within 250 ms while consuming 285 mJ. A k-NN accelerator implemented on Zynqberry platform capable of classifying a touch input of size 4×4 within 26 µs, recording a power consumption of 236 mW has been presented in [13]. For an input touch modality of size $4 \times 4 \times 100$, a 1D-CNN deployed on Arduino Nano BLE 33 managed to offer real-time classification in 26 ms and 128 ms for float and int8 implementations respectively [14].

2.2 Tiny Convolution Neural Network for Touch Modality Classification

Data Structure and Pre-processing. The tactile dataset presented in [8] includes three touch modalities, namely: sliding the finger, brushing a paintbrush, and rolling a washer. Seventy participants were involved in the data collection. Each participant performed both a horizontal and a vertical modality on a 4×4 piezoelectric sensor for 10 s. Samples were obtained at a sampling frequency of 3 kHz and can therefore be represented as a tensor $\phi = 4 \times 4 \times 30,000$. Such input size is impractical for the deployment of existing ML algorithms on hardware platforms [15]. Hence, two pre-processing techniques are employed to reduce the tensor size. First, subsampling is applied to reduce the total activity duration from 10 s to 3.5 s. This is achieved through truncation of silent and noisy intervals. Then, moving average is applied to obtain a tensor $\phi(4 \times 4 \times 8)$. Such tensor size offers a reduced model size with a similar classification accuracy to the original size [16].

CNN Design. The structure of the designed CNN is shown in Fig. 2. The network consists of two convolutional layers, with the second followed by the ReLU activation function, a maxpool layer followed by a ReLU layer, two fully connected layers, with the first followed by a ReLU layer, and the softmax output

layer. The number of layers, and the size of each layer (i.e. number of filters) has been determined based on a best-model approach that offers an acceptable trade-off between complexity and classification accuracy with respect to existing solutions. For the relatively small input dimension 4×4, the adequate filter size used in convolution layers is 3×3. Moreover, in order not to lose the whole input after a convolution layer, convolution layers with padding are adopted. Similarly, only one max pooling layer is used to limit the reduction of the input information. The output layer has three neurons, corresponding to the three touch modalities mentioned above.

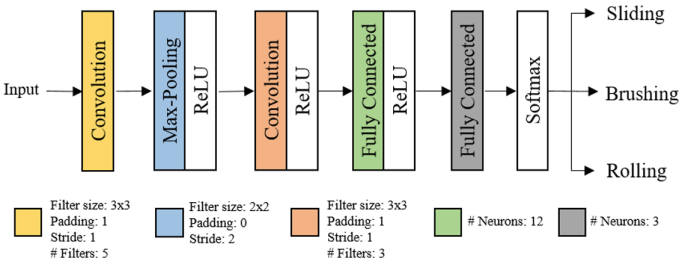


Fig. 2. Proposed TinyCNN architecture.

Training Procedure. The CNN is modeled using TensorFlow and trained for 300 epochs using a batch size of 64. Adam [17] has been chosen as an optimizer with a learning rate of 0.01 and a weight decay of 0.001. These hyperparameters have been obtained through manual fine-tuning. The training has been performed using 5-fold cross validation, and the reported accuracy is the average among the folds; this method is used to be coherent with related works. Training data is normalized. For an input X, the normalized value is obtained as $X_{norm} = (X - \mu)/\sigma$, where μ and σ are the mean and standard deviation on the training set, respectively.

3 Embedded CNN Implementation

3.1 CNN Layers

In order to realize the neural network presented in the previous section, we have implemented a set of components that provide the needed functions for convolutional networks. These components are written in platform-independent C language (i.e., they do not use native OS libraries), which makes them seamlessly employable in most software-programmable edge devices, including but not limiting to microcontrollers. Figure 3 shows how such components are integrated for the different layers' architecture of TinyCNN and how this is optimized for deployment on Cortex-M devices.

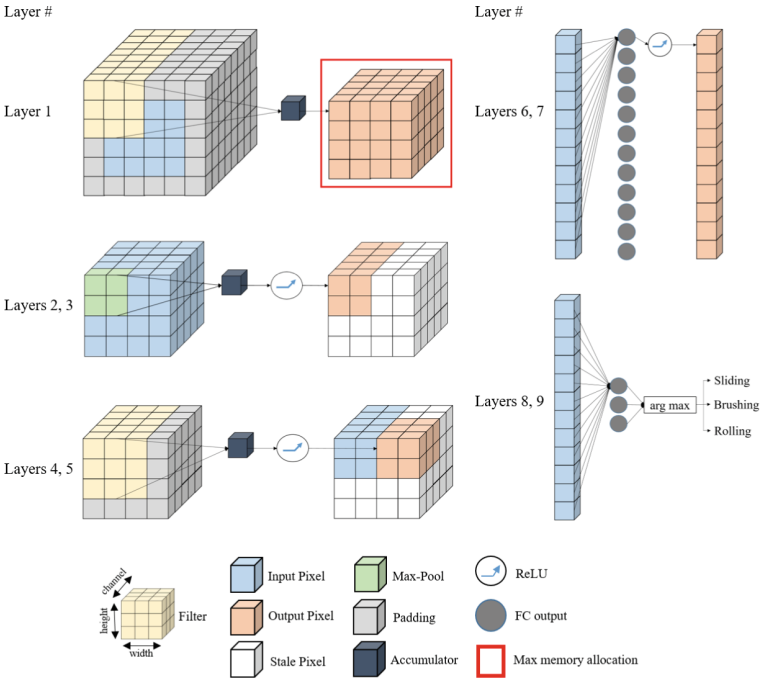


Fig. 3. TinyCNN layers and implementation.

A convolutional component performs 2D floating-point convolution using Multiply Accumulate (MAC) instructions. The convolution of each patch of the image is stored in an accumulator before the results are stored in an output buffer with size equal to the maximum memory buffer requirement, which is pre-computed for the whole network (the block framed in red in Fig. 3). Such memory buffer hosts the values of all the needed neuron activations following a convolutional or a pooling layer, as proposed in [18]. The implementations of the inference phase of neural networks in literature allocate a memory space equal to the sum of the output size of the convolutional layers (e.g., CMSIS-NN [19]). The approach presented in [18], instead, pre-computes and pre-allocates a maximum value of needed memory, without compromising correctness nor performance, under the hypothesis of sequential execution by single-core processors, which corresponds to the vast majority of microcontrollers. The pre-allocated memory block is shown within a red frame in Fig. 3. The technique exploits the fact that convolutions are local operations; the feature maps of a convolutional layer depend only on the input features of that layer. Therefore, after a layer has computed its operations, its memory can be reused to store the features of subsequent layers. The convolutional component supports also layer fusion for the activation function. The activation function is applied directly on the accumulator value before it is written to memory. This (applied, e.g., in layers 4, 5) avoids

the extra reads and writes that would occur if we treated the activation layer as a separate one.

As anticipated, the memory optimization is applied to the max-pooling function as well, which extracts the maximum value in the area it convolves. Our implementation supports in-place computation. The pooling operation is destructive to the input, as each pooling operation results in a single value and allows destroying at least one value, since at least one value will never be used again. This is the founding observation of the memory optimization introduced in [18], which allows directly passing from the fully occupied memory block output by layer 1 to a one quarter occupation after layer 2,3, using only the accumulator as additional temporary memory (i.e., one floating point cell). In the various inference steps, the pre-allocated memory block may feature also some stale (i.e. unused) pixels, as shown in layer 2, 3 and 4,5, because of the lower space's need of some layers. Layer fusion is applied to the output of the max-pooling function as well. The stale pixels resulting from this layer are used to store the output of layer 5, based on the memory replacement strategy proposed in [18].

Finally, the implemented fully connected layer performs vector-to-matrix multiplication using MAC instructions in floating-point representation. Also this layer supports layer fusion for the activation function (layers 6, 7 in Fig. 3). For the inference phase, the softmax function used as the activation function for the multiclass classification problems is replaced by a simpler "argmax" as shown in layers 8, 9. This function operates on a vector and converts every value to zero except the maximum value, where returns 1. This is suitable for the inference phase, that is intended to simply return the only one predicted class, not the probabilities of each possible class.

3.2 Embedded Device and Performance Metrics

In line with the literature on embedded systems for processing tactile data, we assessed the efficiency of the proposed TinyCNN in a deployment on an MCU. In this context, the most relevant metrics are accuracy, area (which, for MCUs, corresponds to the allocated memory space in Flash/RAM), latency, and energy consumption.

Embedded Device. The implemented functions are tested on the STM32 H743ZI2 commercial MCU, hosted on a NUCLEO-144 board. The MCU is an Arm Cortex-M7 core running at 480 MHz, with 512 KB SRAM and 2 MB flash memory.

Latency. Time latency is defined as the time interval it takes the edge device to output a classification decision after the hosted ML model receives an input. We measured the latency on the microcontroller by using the *HAL_GetTick()* function provided by the STM32 hardware abstraction layer's (HAL), which returns the number of elapsed milliseconds.

Energy Consumption. The energy consumption of our implementation is calculated as $E = P \times T$ where P is the power consumption and T is the latency time. The power consumption is calculated as $P = V \times I$ where the average current and voltage values are provided by the STM32CubeIDE.

4 Results and Discussion

Table 1 presents the main performance aspects of the proposed TinyCNN compared to the solutions presented in literature to address the same touch modality classification task. The reported accuracy is derived from related works using the average among 5-folds.The proposed model achieves the highest average classification accuracy of 81.5% among all existing models. Compared to Regularized Least Square (RLS) and SVM in [8], the proposed CNN offers up to 10% accuracy improvement, with a substantial decrease in the number of parameters and FLOPs. Compared to Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU-n; n is the number of neurons) [20], a 17.2% and 26% decrease is achieved in the number of parameters and FLOPs respectively. The reported decrease is with respect to the smallest model (i.e., GRU-10). A 7% accuracy increase is noticed compared to LSTM, which offers the highest accuracy among the solutions presented in [20].

Compared to the 1D-CNN model [14], although the accuracy improvement obtained by our proposed model is less than 2% on average, its size is about 65% smaller for the same input dimensions, which is a key advantage for resource constrained devices.

Table 1. Comparison between different ML algorithms for touch modality classification.

	RLS [8]	SVM [8]	LSTM [20]	GRU-10 [20]	GRU-12 [20]	1D-CNN [14]	Ours
Number of parameters	–	$67.2K$	1113	843	1083	1963	698
kFLOPS	–	$545K$	23.6	17.82	23.6	$530K$[a]	13.2
Average accuracy (%)	73.7	71	74.51	73.1	73.43	79.78	81.5

[a] The reported number of FLOPS includes both pre-processing and inference phases.

The proposed TinyCNN has been deployed on the STM32 H743ZI2 MCU. The TinyCNN requires 2.73 KB for model parameters and 320 bytes for output activations. These requirements permit the allocation of the model to the RAM. This design choice significantly accelerates the inference (compared to Flash allocation), during which an input touch is classified within 0.5 ms, while consuming 65.36 µJ operating under a 3V power source. Although not all ML algorithms listed in Table 1 have been deployed on embedded systems, Table 2

lists the performance of EPUs that have been designed for the same touch modality classification task as ours. As they target different scenarios, each solution differs in terms of ML algorithm, target hardware, pre-processing technique, and expected input size, the comparison is not straightforward. As a common factor, all such EPUs provide real-time classification in less than 400 ms [4].

Table 2. Synopsis of solutions for the touch modality classification task

Reference	ML algorithm	Input size	Pre-processing technique	Hardware target	Time latency	Energy consumption
[12]	SVM	$8 \times 8 \times 20$	Subsampling	Virtex-7	250 ms	285 mJ
		$4 \times 4 \times 20$	Subsampling	PULP	3.3 s	69.3 mJ
[16]	CNN	$4 \times 4 \times 30000$	None	Jetson TX2	75 ms	55.5 mJ
[21]	k-NN	4×4	Time truncation	Zynqberry	25.7 us	6 µJ
[22]	H-CNN	$4 \times 4 \times 8$	Time truncation and subsampling		0.8 ms	42.4 µJ
Ours	TinyCNN			STM32 H743ZI2	0.5 ms	65.36 µJ

Each one of the listed embedded ML algorithms is tailored to the targeted e-skin application. For instance, in a prosthetic application, a person should be able to tolerate the weight of the complete sensory feedback system, thus the adoption of hybrid-CNN or TinyCNN implemented on Zynqberry and MCU, respectively, is more suited, with the final choice based on the trad-off between latency and energy consumption. In an industrial application (e.g. automobile manufacturing), a system with a large number of sensors should be able to process a huge amount of data at once. Hence, higher processing power is required, and an embedded SVM implemented on a Virtex-7 (FPGA), or a CNN implemented on a Jetson GPU, are the main candidates. On the one hand, the latter offers a faster processing time while consuming lower energy. On the other hand the choice could go to the FPGA if a higher value of frames per second in one watt (fps/W) is targeted. Table 2 also shows that an embedded k-NN implemented on Zynqberry offers the fastest classification time and energy consumption. However, k-NN does not create a model during training, and has a linear complexity as the number of training samples increases, which makes it only suitable for small scale applications. Table 2 also mentions pre-processing techniques, as they affect the input dimension, which in turn affects the complexity of the employable ML algorithms. For instance, in SVM, the input tensor is transformed into three matrices, where each matrix undergoes Singular Value Decomposition (SVD). Thus, to reduce the complexity of SVM, it is important to reduce the input size through subsampling. Moreover, there are silent intervals in the tactile data (i.e., regions of zero voltage are detected within the 10 s interval [23]), thus time truncation is applied to remove such regions. In this way, the complexity of the adopted ML algorithm is reduced and the classification accuracy is improved, which leads to reductions in time latency and energy consumption of the embedded system.

5 Conclusion

This paper presents a tiny embedded CNN architecture that we have implemented on a Cortex-M microcontroller for e-skin applications. The tiny CNN features buffer re-use and layer fusion. Compared to existing solutions in the touch modality classification task, the tiny CNN requires a lower number of parameters and FLOPS, while achieving a slight improvement in accuracy. For inference, the embedded TinyCNN achieves real-time classification of tactile data with an energy consumption of 65.36 µJ. We thus argue that our TinyCNN on MCU solution would be advantageous for applications that have area constraints, but still require low time latency and energy consumption. The promising results obtained now suggest the importance of validating the proposed TinyCNN and design decisions in other tactile-based tasks such as object recognition and slip detection.

References

1. Sanchez-Iborra, R., Skarmeta, A.F.: TinyML-enabled frugal smart objects: challenges and opportunities. IEEE Circuits Syst. Mag. **20**(3), 4–18 (2020)
2. Shafique, M., Theocharides, T., Reddy, V.J., Murmann, B.: TinyML: current progress, research challenges, and future roadmap. In: Proceedings - Design Automation Conference, vol. 2021-December, pp. 1303–1306, December 2021
3. Mukherjee, R., Dahiya, R.: Life cycle assessment of energy generating flexible electronic skin. In: 2021 IEEE International Conference on Flexible and Printable Sensors and Systems (FLEPS), Manchester, United Kingdom, pp. 1–4. IEEE, June 2021
4. Johansson, R.S., Flanagan, J.R.: Coding and use of tactile signals from the fingertips in object manipulation tasks. Nat. Rev. Neurosci. **10**, 345–359 (2009)
5. Bhattacharjee, T., Rehg, J.M., Kemp, C.C.: Haptic classification and recognition of objects using a tactile sensing forearm. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, (Vilamoura-Algarve, Portugal), pp. 4090–4097. IEEE, October 2012
6. Kaboli, M., Mittendorfer, P., Hugel, V., Cheng, G.: Humanoids learn object properties from robust tactile feature descriptors via multi-modal artificial skin. In: 2014 IEEE-RAS International Conference on Humanoid Robots, (Madrid, Spain), pp. 187–192. IEEE, November 2014
7. Schill, J., Laaksonen, J., Przybylski, M., Kyrki, V., Asfour, T., Dillmann, R.: Learning continuous grasp stability for a humanoid robot hand based on tactile sensing. In: 2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob), (Rome, Italy), pp. 1901–1906. IEEE, June 2012
8. Gastaldo, P., Pinna, L., Seminara, L., Valle, M., Zunino, R.: Computational intelligence techniques for tactile sensing systems. Sensors **14**, 10952–10976 (2014)
9. Younes, H., Ibrahim, A., Rizk, M., Valle, M.: Data oriented approximate K-nearest neighbor classifier for touch modality recognition. In: 2019 15th Conference on Ph.D Research in Microelectronics and Electronics (PRIME), (Lausanne, Switzerland), pp. 241–244. IEEE, July 2019

10. Alameh, M., Ibrahim, A., Valle, M., Moser, G.: DCNN for tactile sensory data classification based on transfer learning. In: 2019 15th Conference on Ph.D Research in Microelectronics and Electronics (PRIME), (Lausanne, Switzerland), pp. 237–240. IEEE, July 2019
11. Alameh, M., Abbass, Y., Ibrahim, A., Valle, M.: Smart tactile sensing systems based on embedded CNN implementations. Micromachines **11**, 103 (2020)
12. Ibrahim, A., Valle, M.: Real-time embedded machine learning for tensorial tactile data processing. IEEE Trans. Circuits Syst. I Regul. Pap. **65**, 3897–3906 (2018)
13. Younes, H., Ibrahim, A., Rizk, M., Valle, M.: An efficient selection-based kNN architecture for smart embedded hardware accelerators. IEEE Open J. Circuits Syst. **2**, 534–545 (2021)
14. Gianoglio, C., Ragusa, E., Zunino, R., Valle, M.: 1-D convolutional neural networks for touch modalities classification. In: 2021 28th IEEE International Conference on Electronics, Circuits, and Systems (ICECS), (Dubai, United Arab Emirates), pp. 1–6. IEEE, November 2021
15. Gastaldo, P., Pinna, L., Seminara, L., Valle, M., Zunino, R.: A tensor-based pattern-recognition framework for the interpretation of touch modality in artificial skin systems. IEEE Sens. J. **14**, 2216–2225 (2014)
16. Osta, M., et al.: An energy efficient system for touch modality classification in electronic skin applications. In: 2019 IEEE International Symposium on Circuits and Systems (ISCAS), (Sapporo, Japan), pp. 1–4. IEEE, May 2019
17. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 [cs], January 2017
18. Sakr, F., Bellotti, F., Berta, R., Gloria, A.D., Doyle, J.: Memory-efficient CMSIS-NN with replacement strategy. In: Proceedings - 2021 International Conference on Future Internet of Things and Cloud, FiCloud 2021, pp. 299–303, August 2021
19. Lai, L., Suda, N., Chandra, V.: CMSIS-NN: efficient neural network Kernels for arm Cortex-M CPUs, arXiv, vol. abs/1801.06601 (2018)
20. Alameh, M., Abbass, Y., Ibrahim, A., Moser, G., Valle, M.: Touch modality classification using recurrent neural networks. IEEE Sensors J. **21**, 9983–9993 (2021)
21. Ibrahim, A., Younes, H., Alameh, M., Valle, M.: Near sensors computation based on embedded machine learning for electronic skin. Procedia Manufacturing **52**, 295–300 (2020)
22. Younes, H., Ibrahim, A., Rizk, M., Valle, M.: Hybrid fixed-point/binary convolutional neural network accelerator for real-time tactile processing. In: 2021 28th IEEE International Conference on Electronics, Circuits, and Systems (ICECS), (Dubai, United Arab Emirates), pp. 1–5. IEEE, November 2021
23. Younes, H., Ibrahim, A., Rizk, M., Valle, M.: A shallow neural network for real-time embedded machine learning for tensorial tactile data processing. IEEE Trans. Circuits Syst. I **68**, 4232–4244 (2021)

Detecting Ice on Wind Turbine Rotor Blades: Towards Deep Transfer Learning for Image Data

Maria Teresa Alvela Nieto(✉), Hannes Gelbhardt, Jan-Hendrik Ohlendorf, and Klaus-Dieter Thoben

Faculty of Production Engineering, Institute for Integrated Product Development (BIK), University of Bremen, 28359 Bremen, Germany
malvela@uni-bremen.de

Abstract. Wind turbines, in particular, their rotor blades, are not only subjected to specific structural loads but also harsh weather conditions. There exists a risk of ice forming on the leading edge of the rotor blade depending on the location and, notably, at lower temperatures and high humidity. Some of the effects include significant power decreases, turbine damages and shutdowns. Ice detection systems for wind turbines operating in cold climates thus become important. Therefore, this paper proposes a method of ice detection that uses RGB-images, taken from rotating nacelles under different conditions, and pre-trained models of MobileNet, VGG-19 and Xception. The output is an icing prediction that is performed within milliseconds. The novelty of this research lies in utilizing network-based deep transfer learning with unfrozen backbones and learning schedulers. Results showed that the MobileNetV2 obtained up to 99% accuracy. The method outperformed previous research on ice detection by 3% and was evaluated in two different data sets, including near and far views of rotor blades, variety of ice densities and day phases.

Keywords: Wind energy · Ice detection · Deep transfer learning

1 Introduction

Renewable energy expansion is expected to play a key role in climate change mitigation. With a global electricity production share of 27.3% in 2019, renewable energies have established themselves as a major energy source [1]. Wind energy, in particular, contributes to approximately a fifth of all renewable energies [1]. However, ice accretion on wind turbine blades can reduce turbine performance and durability in high altitudes and arctic latitudes, such as in northern Europe, and potentially lead to safety concerns connected with ice shredding [2]. The increasing length of the blades and their modified weight distribution during icing cause high stress on the overall structure, which can enforce a downtime of a wind turbine. Combining a blade heating system with an ice detecting sensor is one solution to the blade icing problem [3]. The sensor, however, can only detect

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 574–582, 2023.
https://doi.org/10.1007/978-3-031-16281-7_54

ice after it has been placed on the rotor blades, which is a drawback of this technology. As a result, there will still be a brief downtime as the heaters have to melt the ice. Besides, ice prediction systems can predict icing before it actually occurs, triggering heaters to prevent ice formation and, in the best-case scenario, completely eliminating turbine downtime [3]. This system, however, consumes far more energy than de-icing [3]. Therefore the installation complexity, quality, and accuracy of existing ice detection systems vary substantially and are far from ideal [2–4]. Still wind farm operators demand both, low-cost effective and plug&play technologies.

This paper proposes deep transfer learning for detecting icing events on image data of rotor blades in near-real time. This novel method relies on pre-trained computer vision algorithms, and the use of a learning rate scheduler [9,10]. The function step decay of the learning rate aims to adjust slowly the unfrozen backbones (base models). To our knowledge, previous research has not attempted to customize the whole backbones for this task [4]. Regarding the detection system, the placement of an RGB-camera on the nacelle, rather than a rotor blade, increases the quality and flexibility of measurements. A RGB-camera and a tiny computer, as a processing unit for inference applications, are low-cost and easily attachable to a turbine. Additionally, operators can visually monitor the actual phase of ice accumulation on predicted images, so that they can better know when to activate or deactivate heating systems. Given that ice accumulation can take up to a few minutes depending on the rotor speed, and ambient temperature and humidity [11], inference times in the millisecond range are not a critical factor in this application. For validation purposes, the presented method is evaluated on a second (new) data set from a wind farm in North America, outperforming previous research by 3% on accuracy.

2 Method

2.1 Data Acquisition and Processing

As input for the models, recorded RGB-images from cameras, installed on the nacelles of wind park turbines, located in the south of Germany, were used. In winter 20/21, images were accessible in 10-seconds intervals for at least two turbines. The original image size is 3840×2160 pixels. As images were taken in the regular turbines operation, neither all images captured a rotor blade nor were taken during day (see Fig. 1). Indeed, images of the same blade(s) can have various backgrounds as a nacelle rotates any time. Night images, which were capture with the help of external lights, were included in the data set. In the labeling process, images were sorted and annotated by hand into the classes 'background', 'rotorblade' and 'ice'. Approximately 4000 images belong to the classes 'background' and 'ice', and 820 images to the class 'rotorblade'. Both classes 'rotorblade' and 'ice' represent rotor blades of the monitored wind turbine. The latter shows additionally icing on rotor blades at each phase of the ice formation (ice density). The 'background' class includes all other images, e.g. far views of wind turbines. Figure 1 shows exemplary images for the classes

Fig. 1. Exemplary images from the data set: 'Background' (left, rotor blades are only in backdrop), 'rotorblade' (middle, a rotor blade with no ice) and 'ice' (right, a rotor blade with ice).

of the data set. The data set is not publicly available since a private research partner delivered it.

As the provided data set has a large class imbalance (10%-45%-45%), the class 'rotorblade' was augmented for the training process using a Gaussian function and an unsharpening filter. In this sense, the number of samples for each category was approximately the same as the training data. To artificially enhance the size of the data, a random height and width shift up to a fraction of 10% was applied to each picture dimension. Afterwards, the images were resized using the pre-processing function corresponding to each selected model.

2.2 Model Architecture and Implementation

As base models, the VGG19 [5], MobileNetV2 [6] and the Xception [7] networks were investigated. This study used the weights from ImageNet [8]. In contrast to previous research (e.g. [4]), this method investigated the ability to adapt slowly the learned features from ImageNet to our target domain (blade icing events) by using unfrozen backbones. The transfer learning method used in this paper can be categorized as network-based deep transfer learning [9]. Except for the output, a fully connected layer, no additional layers were attached to the base models (unfrozen backbones). Therefore, new features from the target domain

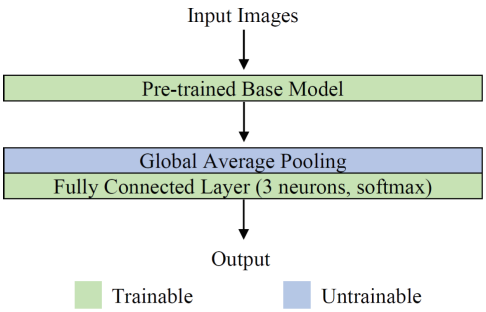


Fig. 2. The model architecture shows the trainable base model, a global average pooling function for the semantic part, and the fully connected layer attached to it.

data set were extracted, and then the classification task was handled by the newly added layer. Figure 2 shows the model architecture of this method.

The maximum number of epochs for each model and training run is limited to 15. An SGD optimizer and a batch size of 16 were utilized for the training. Regarding the learning rate scheduler, the training process was initiated with a learning rate of 1.5e−03, and the momentum of the SGD optimizer was set to 0.9. During training, the scheduler sought to adjust the learning process by reducing the learning rate by a factor of 0.94 each second epoch, according to a step decay scheduler. After ten cycles, the learning rate was further reduced to 1.5e−04 while the scheduler factor did not change. For the experiments, the practical implementation of the models relied on Python libraries, including the framework Keras with Tensorflow as backend. The code is made publicly available on GitHub at https://github.com/malvela/WindTurbine-IceDetection. All training processes were performed on 2 x Nvidia GTX 1080 TI.

For each split of a three repetition 5-fold cross-validation, the pre-processed data were split into 60% training, 20% validation, and 20% testing, resulting in 15 training and evaluation phases per model configuration. The model performance of the validation and test splits was measured by the metrics accuracy and F1-score, as described in Table 1. Confusion matrices were also calculated to assess performance and make it easier to study true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) with regard to each class.

Table 1. Used metrics for the evaluation and validation phase.

Accuracy	Precision	Recall	F1-score
$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$\frac{2}{Precision^{-1}+Recall^{-1}}$

3 Results and Discussion

Table 2 shows the results on average of the test data set. For all models, the accuracy is higher than the F1-Score. A reason for this effect can be the class imbalance in the test data, as the class 'rotorblade' for evaluation and validation was not augmented (see Fig. 3). Moreover, the results of MobileNetV2 and Xception and their confidence interval are very close to each other (considering rounding up to two decimal places). Instead, VGG19 model performed worse than the other two evaluated models for this task, which confirms previous works in different domains [7,12]. For recent task domain efforts, such as [4], their VGG models achieved the best performance while MobileNetV2 obtained 0.56% (with interval confidence of 0.21). In contrast to previous work, 3 out of the 15 VGG19 runs on our data set presented an F1-score value lower than 0.21. For the remaining training runs, the model scored above 0.9. This effect is reflected by the high confidence interval of the VGG19 shown in Table 2.

Besides precision metrics, inference times can be regarded as another critical attribute in evaluating models, especially when production-ready models have

to either process a large amount of data in a limited time or are a crucial part of a live-monitoring system. According to Table 2, VGG19 is making predictions on our images even twice faster than the other two models. The three models were able to predict at least 30 images per second on the used GPUs. Inference times in this range are excellent for blade icing detection tasks, as water freezing processes take much more time than a few seconds [11].

Table 2. Average accuracy, F1-score and inference time (in seconds) on the test data as a result of a five fold 3-repetition cross validation. For both accuracy and F1-score twice the standard deviation is stated. The highest scores are in bold. Inference time in seconds.

Model	Accuracy	F1-score	Inference time
MobileNetV2	**0.95** $\pm$ 0.01	**0.92** $\pm$ 0.02	0.028
VGG19	0.83 $\pm$ 0.17	0.78 $\pm$ 0.22	0.012
Xception	**0.95** $\pm$ 0.01	**0.92** $\pm$ 0.02	0.027

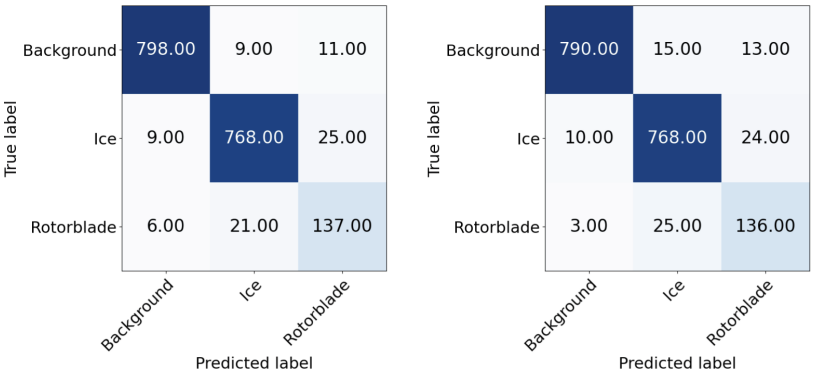


Fig. 3. Confusion matrices that show the absolute number of true and predicted classes for MobileNetV2 (left) and Xception (right) from the test data.

The rows of the confusion matrices in Fig. 3 represent the number of images in an actual class, while each column represents the predicted images assigned to each class. The two models used the same set of test images, which belong to various turbines and ice formation stages. From the model predictions, it can be assumed that the performance of MobileNetV2 and Xception is very similar. However, the fraction of misclassified samples for the 'background' class is significantly lower than the fraction of the other two classes. This can indicate that the model clearly distinguishes between 'Background' and the other classes. In other words, a bit higher fraction of samples belonging to the classes 'ice' and 'rotorblade' is classified wrongly. The diagonal line develops darker as more images are successfully allocated to the true class.

As a proof of method applicability, a new data set belonging to a wind park located in North America and earlier introduced by [4] was analyzed. The model accuracy on the new data has increased compared to the prior data collection. The average scores are summarized in Table 3. For the implementation, we used approximately the same number of images analyzed by [4]. The re-sampling method, training iterations as well as the data split (60%-20%-20%) and augmentation process of the training data set were identical to the previous experiment (see Table 2).

Table 3. Mean results of applying the presented method in a new data set; twice the standard deviation is assigned (best scores in bold). Inference time is in seconds.

Model	Accuracy	F1-score	Inference time
MobileNetV2	**0.99** $\pm$ 0.01	**0.99** $\pm$ 0.01	0.013
VGG19	0.86 $\pm$ 0.14	0.83 $\pm$ 0.17	0.007
Xception	0.98 $\pm$ 0.01	**0.99** $\pm$ 0.01	0.015

The values reported in Table 3 exhibit a similar pattern to those summarized in Table 2. At this time, the MobileNetV2 and Xception networks outperformed significantly the VGG19. Both models achieved scores of 0.99%. Compared to the previous data set, the VGG19 improved its F1-score by 5%. Curiously, 3 out of 15 cross-validation splits had, again, shallow values (around 0.55) for the VGG19. The remaining 12 splits, all got values greater than 0.96. If the total number of iterations is reduced to the ten best outcomes, the VGG19 model can provide results comparable to those of the MobileNetV2 and Xception.

The confusion matrices in Fig. 4 present the prediction of on average MobileNetV2 and Xception results on the second data set. Even though the 'ice' class is the one with fewer samples (42 of 439 images), only MobileNetV2 assigned wrongly 'rotorblade' to a 'ice' image. The overall wrong classification for both models is 5 out of 439 images. In both models, the classes 'background' and 'rotorblade' were not clearly differentiated. One would expect to rise up the prediction accuracy if these two classes are treated as a single class. The treatment of one instead of two classes is very specific to the application. A unique class dedicated to only rotor blade images for optical inspection of blade corrosion, for example, would be of interest. A model that first separate images with rotor blades from no rotor blade images, followed by a second model to split blade icing from blade images, may be considered as a future work.

The results stated in Table 2 and Fig. 3 show the ability of the method to predict ice on rotor blade images accurately. The evaluation of the method on a second data set, which stated more qualitative pieces of information, outperformed previous outcomes and research work. The use of unfrozen backbones and learning rate scheduler on the second data set outperforms previous methods by a margin of +3%. Concerning the poor prediction quality of the VGG19, one might speculate that an unfavorable test set could have led to a relatively

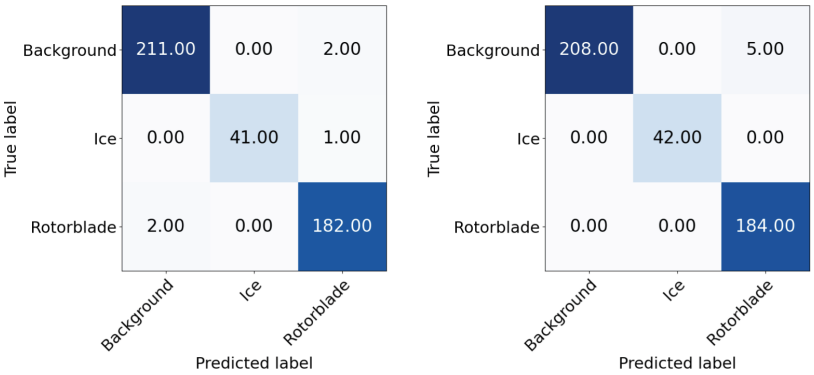


Fig. 4. Confusion matrices showing the absolute number of images of the MobileNetV2 (left) and Xception network (right) predicting the new test data used in [4].

high gradient update, overwriting features of the pre-trained models, and failing into wrong interpretations. This fact may suggest that the initial learning rate is a critical attribute of the method on the VGG19 network. Therefore, further experiments with lower initial learning rates for training a VGG19 model will be investigated in a future.

The computational complexity of the proposed method is summarized in Table 4. From this table, the VGG19 and Xception model had almost comparable number of (trainable-) parameters, while the MobileNetV2 only required about $\frac{1}{10}$ of that number. The number of trainable parameters is remarkably higher than other research works, for instance [4], as the backbones of this method were unfrozen. The obvious benefit of having many parameters is that the representation can be much more complicated functions than with fewer parameters. Early stopping or stronger regularization techniques can prevent overfitting in these cases. The results in Table 3 thus showed that the models outperformed previous studies. Beyond that, however, the role of trainable parameters for the deployment on resource-constrained devices is relevant in terms of memory and time. Table 5 represents how practicing model quantization to make a smaller model that can run under the constraints of a Jetson Nano 4 GB without compromising on accuracy is possible. The model weights were recorded as 32-bits floating-points numbers in the original weight format (trained weights), which is a subset of the hierarchical data format (h5). Using a trt-converter, these weights were quantized to 16 bits. The models conversion led to a halved model size (see Table 5). The stated inference times in Table 5 include reading, preprocessing, predicting and writing predictions in the simulated production environment. For MobileNetV2 and Xception the inference times were successfully reduced. The conversion of the VGG19 model did not improve the inference time at all. The accuracy was measured on the test set for one cross-validation fold and was not compromised, retaining the values suggested in Table 3.

Table 4. Total and trainable parameters of each model and default resolution of input image.

Model	Total parameters	Trainable parameters	Input image size
MobileNetV2	$2,261,827$	$2,227,715$	224×224
VGG19	$20,025,923$	$20,025,923$	224×224
Xception	$20,867,627$	$20,813,099$	299×299

Table 5. Model comparison with regard to size (in megabytes) and inference time (in seconds) using quantization. All results were produced on a Jetson Nano 4 GB.

Model	File size	Inference time
MobileNetV2_h5	18.504	0.116
MobileNetV2_fp16	9.123	0.057
VGG19_h5	161.634	0.338
VGG19_fp16	80.836	0.339
Xception_h5	171.672	0.324
Xception_fp16	85.734	0.232

Data sparsity and fluctuating rotor blade characteristics (shape, length, color code-lines at the tip,..) could drive future research to further investigate the knowledge transfer gained on an existing wind park with specific rotor blades to new wind park installations. One limitation of the presented method is the use of a RGB-camera. Under extreme weather conditions, e.g., fog or snow, the images become too blurry that rotor blades are not visible. Future work should consider using thermal cameras, which can still capture actual blade information under harsh weather.

4 Conclusions

This paper presents a method for detecting icing on wind turbine rotor blades as a first step to improving wind turbines efficiency at low-cost. The proposed method is based on deep transfer learning and uses convolutional neural networks that accept RGB-image data as inputs, and it outputs icing predictions within milliseconds in an inference application. The method was proved in two data sets from two different wind parks. In our test setup with a Jeson Nano 4 GB, MobileNetV2 had the best performance, with accuracy and F1-score of both 99%, an inference time of 57 ms and a size of 9 megabytes. The model performance gap along data sets demonstrates the importance of a good camera, high-quality and well-labelled images. Since data acquisition and labeling are both costly and time-consuming, future work should focus on creating synthetic

data, generalizing task-domain models and developing novel transfer mechanisms for upcoming rotor blade characteristics.

Acknowledgments. This work is part of the research project "EisAuge - Ice detection on wind turbines using AI-assisted image processing", funded by the European Regional Development Fund, funding code VE0126C. We thank wpd windmanager GmbH Co.KG the provision of the datasets and the collaboration in this research.

References

1. Renewables 2020 Global Status Report. https://www.ren21.net/wp-content/uploads/2019/05/gsr_2020_full_report_en.pdf
2. Pryor, S.C., Barthelmie, R.J.: Climate change impacts on wind energy: a review. Renew. Sustain. Energy Rev. **14**, 430–437 (2010). https://doi.org/10.1016/j.rser.2009.07.028
3. Parent, O., Ilinca, A.: Anti-icing and de-icing techniques for wind turbines: critical review. Cold Regions Sci. Technol. **65**, 88–96 (2011). https://doi.org/10.1016/j.coldregions.2010.01
4. Kreutz, M., Alla, A.A., Eisenstadt, A., Freitag, M.J., Thoben, K.D.: Ice detection on rotor blades of wind turbines using RGB images and convolutional neural networks. Procedia CIRP **93**, 1292–1297 (2020)
5. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations (2015). https://doi.org/10.48550/arXiv.1409.1556
6. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: MobileNetV2: inverted residuals and linear bottlenecks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018). https://doi.org/10.48550/arXiv.1801.04381
7. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807 (2017). https://doi.org/10.48550/arXiv.1610.02357
8. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
9. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A survey on deep transfer learning. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds.) ICANN 2018. LNCS, vol. 11141, pp. 270–279. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01424-7_27
10. Schaul, T., Zhang, S., LeCun, Y.: No more pesky learning rates. In: Proceedings of the 30th International Conference on International Conference on Machine Learning, vol. 28 (2013). https://doi.org/10.5555/3042817.3042975
11. Wei, K., Yang, Y., Zuo, H., Zhong, D.: A review on ice detection technology and ice elimination technology for wind turbine. Wind Energy **23**(3), 433–457 (2020). https://doi.org/10.1002/we.2427
12. Canziani, A., Paszke, A., Culurciello, E.: An analysis of deep neural network models for practical applications. arXiv (2016). https://doi.org/10.48550/arXiv.1605

Towards Real-Time Human Detection in Maritime Environment Using Embedded Deep Learning

Mostafa Rizk[1,3,4(✉)], Fatima Slim[3], Amer Baghdadi[1], and Jean-Philippe Diguet[2]

[1] IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238 Brest, France
mostfa.rizk@imt-atlantique.fr
[2] CNRS, IRL CROSSING, Adelaide, Australia
[3] Physics and Electronics Department, Lebanese University, Beirut, Lebanon
[4] CCE Department, Lebanese International University, Beirut, Lebanon

Abstract. Marine search and rescue missions necessitate a lot of effort and expenses. The use of technological advancements facilitates discovering and locating individuals and aids in the directing of rescuers and medical teams. This has the potential to save human lives while also lowering costs. The characteristics of the marine environment create additional challenges for computer vision techniques used to detect the presence of human in a scene. Currently, artificial intelligence (AI) techniques based on convolution neural networks (CNNs) provide solid solutions to detect and locate objects. In this paper, the relevance of the emergent You Only Look Once (YOLO) in detecting humans in maritime environment is investigated. The available models of YOLOv4 are trained using a custom dataset. The trained models are evaluated using recognized evaluation parameters. In addition, the inference speed is reported targeting embedded low-power hardware platforms dedicated for AI applications.

Keywords: Deep learning · YOLO · Maritime · Human detection · Man overboard · Search and rescue

1 Introduction

Maritime search and rescue (SAR) missions are crucial for most coastal states. According to the International Organization for Migration 218,062 irregular maritime migration attempts are recorded in the Mediterranean Since 2014 [1] From which, 23,939 dead and missing persons are recorded during attempted overseas crossings. Furthermore, the European Maritime Safety Agency reports in the Annual Overview of Marine Casualties and Incidents 2021 [2] that during the 2014–2020 period, 367 marine casualties resulted in a total of 550 lives lost and

This work was supported in part by the Regional Council of Bretagne through the ODESSA FEDER project.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 583–593, 2023.
https://doi.org/10.1007/978-3-031-16281-7_55

6921 injuries in the waters of European Union (EU) Member States or involving EU ships. The ability to quickly locate missing people aids in the direction of rescuers and medical personnel, which plays an important role in increasing the chances of saving human lives while also lowering costs.

Years ago, visual surveillance in the maritime domain has been explored. However, most surveillance activities have been assigned to areas near the coasts and ports and mainly depend on human monitoring and analysis for security reasons. Computer vision techniques are also adopted in few works. However, videos and images capturing maritime environment pose challenges that are absent or less severe in other environments such as the dynamic nature of the background, unavailability of static cues, presence of small objects at distant backgrounds and illumination effects [3]. These challenges impact the efficiency of traditional computer vision techniques in detecting individuals in marine environments.

Recently, deep learning approaches have introduced efficient solutions to detect, classify and localize several objects in images and videos. In particular, the evolution of neural networks architectures has elevated the performance to a point that they are considered on par with human performance for some of these problems. However, the detection performance comes at the cost of increased hardware resources and power consumption especially for real-time scenarios with high requirements of accuracy and precision. You Only Look Once (YOLO) has been recently introduced as an efficient unified model of all phases of a CNN for doing object detection in real-time. The recent version of YOLO, so called YOLOv4, has been justified to detect objects in real-time with high level of precision. Several models of YOLOv4 exist, with different architecture specifications and consequently different detection performance in terms of accuracy and precision, detection speed and required energy budget.

The growing use of artificial intelligence (AI) based detection methods is of great interest in aiding SAR missions [4–8]. However, only few works have addressed the detection of humans in open water or for man overboard accidents [9–11]. Other available works adopting deep learning in marine environment have focused mainly on the detection of sea ships [12]. This work aims to enable efficient detection and localizing of floating humans in real-time based on AI techniques. In particular, the relevancy of YOLOv4 [13] in detecting humans in maritime environments to aid marine SAR missions is addressed. The work includes collecting a custom dataset, training different YOLOv4 available models and evaluating the trained model using mean average precision (mAP), precision, recall, and F1-score. Also, the trained models are implemented targeting Jestson Nano and Jetson Xavier development kits from Nvidia. For different power modes, the inference speed is attained while processing real-life videos. The obtained results show that YOLOv4 can achieve real-time detection when implemented on low-cost, small size embedded platforms with reduced power consumption. This paves the way to develop airborne systems or edge embedded systems mounted on shore, moving boats or floating buoys that can be exploited to facilitate search and rescue missions and in optimizing the man overboard signaling systems.

2 Background

2.1 Object Detection

Previously, object detection has been achieved using computer vision techniques based on feature extraction such as histogram of oriented gradients (HOG) [14] and scale-invariant feature transform (SIFT) [15]. Currently, artificial intelligence (AI) techniques based on convolution neural networks (CNNs) are the dominant methods for object detection, which compromise both classification and localizing of objects within the image by determining bounding boxes (coordinates and size) around the objects of interest. Several techniques based on CNN are developed targeting object detection. Two-stage models such as region-convolutional neural network (R-CNN) [16] apply classification of objects based on pre-selected regions. The post-processing operations required to refine the bounding boxes, eliminate duplicates and adjust the detection scores increase the complexity and impact the speed of detection. Despite the introduction of R-CNN enhanced versions [17,18], real-time detection has not been granted.

You Only Look Once (YOLO) has been proposed in [19] as an efficient unified model of all phases of a CNN for doing object detection in real-time. Several versions of YOLO have been developed by modifying the network architecture. In YOLOv2 [20], the fully connected layers at the end have been eliminated and Darknet-19 architecture has been adopted. YOLOv3 [21] uses Darknet-53 architecture and inherits the concept of residual networks. The detections are made at 3 different scales which enables the detection of small objects. Recently, YOLOv4 [13] object detection method has been introduced. It outperforms other available methods in terms of speed and accuracy performance. The experiments targeting Microsoft Common object in context (COCO) dataset [22] show that YOLOv4 is faster and more accurate than real-time neural networks EfficientDet [23] and RetinaNet [24] provided by Google and Facebook respectively.

The architecture of YOLOv4 consists of the backbone, neck and dense prediction so-called the head. The backbone is in charge of extracting features. The neck aggregates the features and delivers them to the detection head. Based on several experiments and comparisons [13], CSPDarknet53 is selected for the backbone. Spatial Pyramid Pooling block (SPP) is added to the PANet path-aggregation neck. The anchor based YOLOv3 is adopted as detection head in YOLOv4.

YOLOv4 exploits a set of universal methods that are assumed to improve CNN accuracy for majority of models, tasks, and datasets. These universal methods are data augmentation (DA), Weighted-Residual-Connections (WRC), Cross-Stage-Partial-connections (CSP), Cross mini-Batch Normalization (CmBN), Self-adversarial-training (SAT) and Mish-activation. These universal methods are implemented in combination with new devised methods such DropBlock regularization, and Complete-IoU loss. YOLOv4 employs these available methods in two ways in order to create a more efficient and powerful object detection model: *Bag-of-Freebies* and *Bag-of-Specials*. *Bag-of-Freebies* compromises training strategies and pre-processing methods. Adopting these strategies

enhance the training without impacting the inference performance as training is done offline. Data augmentation is used to alleviate the degree of variability of training images in order to increase the robustness of the detection during inference against unknown environments. Data augmentation includes pixel-wise computer vision techniques such cutmix, mosaic, image resizing, blurring, image rotating, random scaling, flipping, cropping and changing the exposure, saturation and hue. Focal loss is also adopted to address the issue of data imbalance existing between various classes. Label smoothing is used to convert hard labels into soft labels leading to improving the robustness of the model. *Bag-of-Specials* contains architecture-related plug-in modules and post-processing methods introduced. Mish activation is used for both backbone and detector. CSP and Multi-input weighted residual connections are selected for the backbone. SPP-block, SAM-block, PAN path-aggregation block are added to the neck/detector.

2.2 Human Detection Using Deep Learning Methods

Several works have adopted deep learning techniques to detect individuals for several applications such as social distancing [25], crowd detection, security and search and rescue missions [4–8]. However, few works have addressed human detection in marine environment. In [9], the authors have exploited YOLOv3 Tiny to detect human swimming in open water via areal images. The authors have deployed the trained network on NVIDIA Jetson TX1 platform to enable real-time detection of human in search and rescue missions using UAVs. In [26], SSD and YOLOv3 have been examined to detect man overboard event detection. The authors have not presented the performance results. In [10], Faster R-CNN has been employed to locate the person in water using thermal images. In [11], YOLOv3 has been utilized to detect and localize human in marine environment using images captured by UAVs for search and rescue missions. The authors have focused on analyzing the effects of flight altitude on the detection performance. The used dataset for training, validating and testing includes 450 images only, which are collected in one location. Note that in [10] and [11] the training and testing results in terms of precession are only shown without presenting the achieved detection speed or indicating the used target device.

3 Method

3.1 Dataset

We create a diverse dataset of images showing humans in maritime environment. The images are collected from several internet resources. We make use of the dataset published by [9]. The dataset offers images extracted by videos captured by the means of UAV for Humans swimming in open water. We edited this dataset by eliminating images with high similarity. Also, a great effort is done to enhance the labeling by adjusting the existing bounding boxes to meet with the dimensions of the persons and by adding bounding boxes of unlabeled persons. In addition,

we add 2000 new images including showing persons in maritime environment with different positions and from different perspectives. The number of humans in the scene varies between the gathered images. In addition, the images show human bodies in numerous positions and different perspectives and scales, and have various backgrounds, lighting conditions and resolutions. The final dataset includes 6462 images with 16795 bounding boxes[1]. The images are split randomly by 70% as training dataset, 10% as validation dataset and 20% as testing dataset. Table 1 shows the distribution of images and objects in each dataset.

Table 1. Specifications of the created dataset

Dataset	Training	Validation	Testing	Total
Number of images	4463	666	1333	6462
Number of objects	11913	1677	3205	16795

3.2 Target Models

In this work we examine three different YOLOv4 networks: YOLOv4 Large, YOLOv4 Tiny and YOLOv4 Tiny-3l. The original YOLOv4 network consists of 162 layers and uses mish activation functions. YOLOv4 Tiny is the compressed version of YOLOv4. It uses the simplified network structure of CSPDarknet53-tiny. It compromise 38 layers with LeakyRelu activation functions and only two detector heads. YOLOv4 Tiny-3l architecture is similar to YOLOv4 Tiny, but with three detector heads. Table 2 presents the target networks specifications.

Table 2. Specifications of the targeted YOLOv4 models

Model	Number of layers	Activation function	Model weights' volume (MB)
YOLOv4	162	Mish	256.2
YOLOv4 Tiny	38	LeakyRelu	23.5
YOLOv4 Tiny-3l	45	LeakyRelu	24.5

3.3 Training

The training is conducted using the Darknet framework [27] using Quadro RTX 4000 from Nvidia. Transfer learning is adopted in order to maintain the generalization. We make use of the weights generated in previous training processes of networks with similar architecture specifications targeting COCO dataset. Note that the imported weights of the feature extraction layers are kept; whereas, the weights of the neck and the detector layers are eliminated. The networks' general

[1] https://www.kaggle.com/datasets/mostafarizk/maritimesar.

architectures have not been altered. Only the depth size of the three convolution layers allocated before the YOLO detector layers are adjusted. The number of filters in these three convolution layers are modified considering our case where only one class (Person) is targeted.

The number of images per batch is set to 64. The total number of iterations is set to 2000. The initial learning rate for training is set to 0.001 and it scales down two times by 0.1 at iteration 1600 and 1800. The input images are down sampled into 416×416 or 608×608. While training the models, data augmentation is activated. Mosaic data augmentation type is used where 4 images are merged into one. When activated, Cutmix data augmentation type is applied for the classifier only. The saturation of input images and their exposure (brightness) are randomly changed as well as the rotation.

The models are validated using the validation dataset. Mean average precisions (mAP) is calculated during training for each 4 epochs. Figure 1 illustrates the training performances. Note that the blue curves correspond to the training losses whereas the red curves corresponds to the computed mAP values. The mAP calculation starts after 1000 iterations and it adopts the AP50 metric defined in the MS COCO competition (same to the metric of precision in the Pascal VOC competition) and uses the following expressions to compute the Precision and Recall values:

$$P = \frac{TP}{(TP + FP)} \qquad R = \frac{TP}{(TP + FN)} \tag{1}$$

where P is the Precision, R is the Recall and TP, FP and FN stand for True Positive, False Positive and False Negative respectively. Table 3 shows the required time for training the targeted networks with different input resolutions using Nvidia Quadro RTX 4000.

3.4 Evaluation

The trained models are evaluated using the test dataset. Sample detection results from network testing are shown in Fig. 2. The figure shows that trained models are able to accurately detect and classify the presence of human bodies in different maritime environments. Table 3 shows the obtained mean average precision considering VOC07 and VOC12 performance metrics [28]. In addition, the table shows the obtained values of precession, recall, F1-score and average intersection over union (IOU) considering 0.5 IOU threshold.

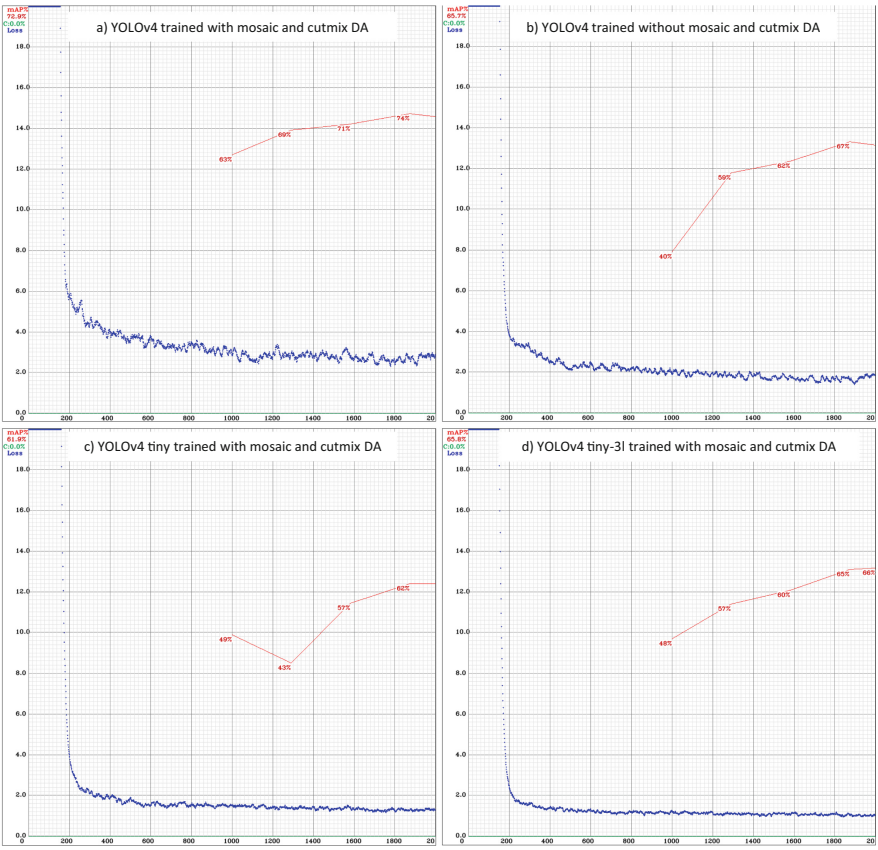


Fig. 1. Sample training performances

Furthermore, the inference speed of trained models is evaluated using several captured videos targeting embedded platforms. Table 4 shows the obtained speed of the trained models in frames per second (FPS) when applied to the captured videos on Jetson Nano and Jetson Xavier NX development kits while operating on different power modes. Both used kits are small powerful computers that allow running neural networks for applications like image classification, object detection, segmentation, etc. Jetson Nano provides 472 GFLOPS of FP16 computing performance with 5W and 10W of power consumption. Whereas, Jetson Xavier NX provides up to 21 TeraOPS of compute performance in configurable

Table 3. Evaluation results of the trained YOLOv4 models

Target model	Image resolution	Training time (h)	Data augmentation	mAP VOC07	mAP VOC12	Precession	Recall	F1 score	avg IOU
YOLOv4 Large	416 × 416	03:22	–	60.46	58.78	61.61	70.80	65.88	62.14
			mosaic	64.27	65.66	62.57	75.48	68.42	63.8
			mosaic+cutmix	65.63	69.04	61.95	78.03	69.07	64.83
	608 × 608	06:00	–	55.16	59.15	66.15	69.98	68.01	63.13
			mosaic	64.64	64.91	66.74	73.45	69.93	64.92
			mosaic+cutmix	65.82	69.37	63.96	78.28	70.4	65.39
YOLOv4 Tiny	416 × 416	00:24	–	57.00	56.53	49.03	73.39	58.79	61.34
			mosaic	56.34	56.10	48.25	73.95	58.40	61.75
			mosaic+cutmix	56.90	56.90	47.40	73.85	57.74	61.85
	608 × 608	00:35	–	59.29	60.07	53.98	74.91	62.75	62.08
			mosaic	60.91	63.10	52.83	77.00	62.66	62.58
			mosaic+cutmix	60.59	62.47	53.04	76.57	62.67	62.66
YOLOv4 Tiny-3l	416 × 416	00:25	–	54.89	53.31	53.78	72.32	61.69	60.97
			mosaic	55.28	54.17	53.16	73.04	61.53	61.57
			mosaic+cutmix	55.80	55.41	53.09	73.95	61.81	61.88
	608 × 608	00:47	–	59.12	57.46	59.27	70.92	64.57	61.81
			mosaic	59.54	59.43	56.63	73.26	63.88	62.35
			mosaic+cutmix	60.08	59.92	57.60	73.17	64.46	62.21

Table 4. Average detection performance in FPS

Trained model	Input frame resolution	Jetson Nano		Jetson Xavier NX					
		Mode0	Mode1	Mode0	Mode1	Mode2	Mode3	Mode4	Mode5
		10W	5W	15W 2CORE	15W 4CORE	15W 6CORE	10W 2CORE	10W 4CORE	10W Desktop
YOLOv4 Large	416 × 416	2.0	1.5	10.1	10.7	10.8	8.8	9.4	6.5
	608 × 608	1.0	0.7	5.6	5.7	5.7	5.0	4.9	3.7
YOLOv4 Tiny	416 × 416	19.2	12.5	58.4	80.6	72.0	54.3	67.4	53.0
	608 × 608	9.6	6.5	35.3	44.3	45.6	31.3	39.4	30.3
YOLOv4 Tiny-3l	416 × 416	16.8	10.9	50.1	60.0	69.9	43.6	58.6	48.0
	608 × 608	8.5	5.7	32.6	39.7	40.3	35.1	35.7	26.8

10W or 15W power budgets by capping the GPU and CPU frequencies and the number of online CPU cores at a pre-defined level. Figure 3 shows samples of the obtained detection results in captured video sequences. The obtained results show that applying DA enhances the detection performance (mAP, precision, recall, F1-score and average IOU). The use of cutmix DA increases the enhancement ratio in most of the cases. The use of higher image resolution enhances the mAP performance but at the cost of reduced inference speed and longer training time.

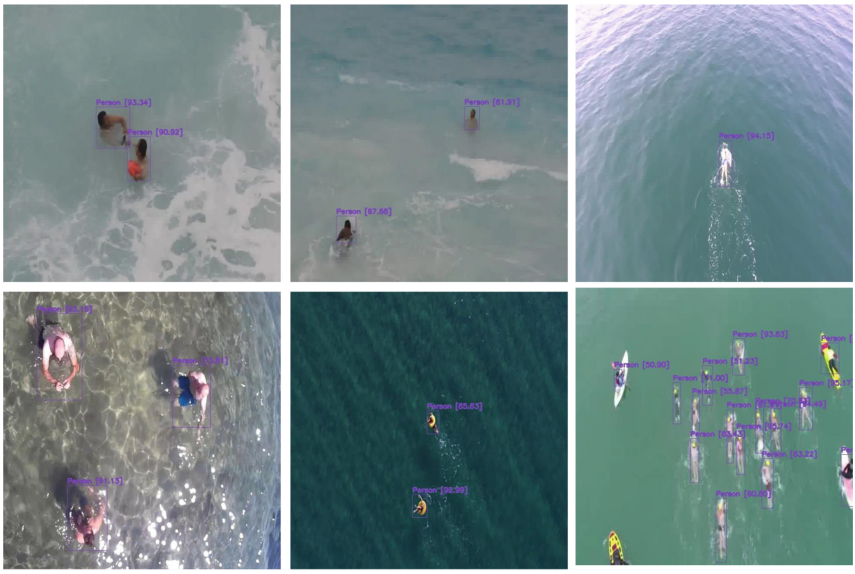


Fig. 2. Sample detection results in testing dataset images

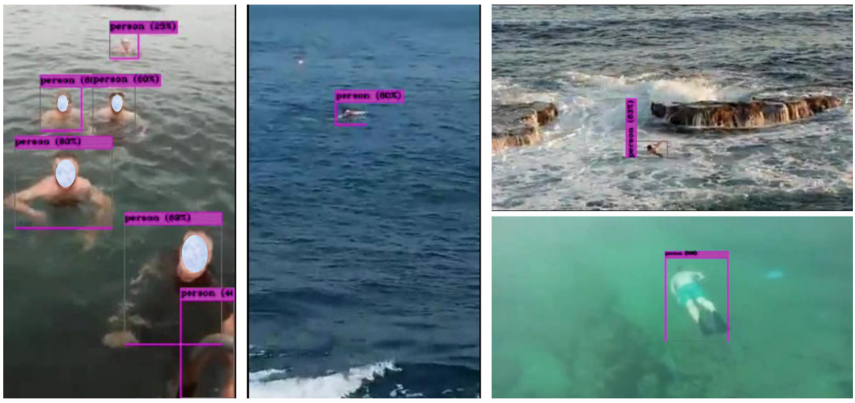


Fig. 3. Samples of the obtained detection results in video sequences

4 Conclusion

In this paper, the use of YOLOv4 in detection of humans in maritime environments is investigated. Available YOLOv4 architectures are trained on a custom dataset. The trained models are evaluated in terms of mAP, precession, recall and average IOU. Also, the performances of the models are examined on embedded platforms using our own videos showing humans in open water. The obtained results show that YOLOv4 can achieve real-time detection of humans in maritime environment with acceptable accuracy and precession. For example,

YOLOv4 Tiny achieves an inference speed of 45.6 FPS with mAP of 63.10 when running on Jetson Xavier NX considering 608×608 resolution. Future work will include applying optimization techniques such as quantization and pruning to increase the inference speed and study their impact on the detection performance.

References

1. International Organization for Migration Missing Migrants Project website. https://missingmigrants.iom.int. Accessed 1 May 2022
2. E. M. S. Agency: Annual overview of marine casualties and incidents 2021, EMSA, Annual Report, December 2021
3. Prasad, D.K., et al.: Challenges in video based object detection in maritime scenario using computer vision. arXiv preprint arXiv:1608.01079 (2016)
4. Castellano, G., Castiello, C., Mencar, C., Vessio, G.: Preliminary evaluation of TinyYOLO on a new dataset for search-and-rescue with drones. In: International Conference on Soft Computing Machine Intelligence (ISCMI), pp. 163–166 (2020)
5. Liu, C., Szirányi, T.: Real-time human detection and gesture recognition for on-board UAV rescue. Sensors **21**(6), 2180 (2021)
6. Rizk, M., Slim, F., Charara, J.: Toward AI-assisted UAV for human detection in search and rescue missions. In: 2021 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, December 2021, pp. 781–786 (2021)
7. Sambolek, S., Ivasic-Kos, M.: Automatic person detection in search and rescue operations using deep CNN detectors. IEEE Access **9**, 37 905–37 922 (2021)
8. Rosero, R.L., Grilo, C., Silva, C.: Deep learning with real-time inference for human detection in search and rescue. In: Abraham, A., Piuri, V., Gandhi, N., Siarry, P., Kaklauskas, A., Madureira, A. (eds.) Intelligent Systems Design and Applications, pp. 247–257. Springer, Cham (2021)
9. Lygouras, E., et al.: Unsupervised human detection with an embedded vision system on a fully autonomous UAV for search and rescue operations. Sensors **19**(16), 3542 (2019)
10. Feraru, V.A., Andersen, R.E., Boukas, E.: Towards an autonomous UAV-based system to assist search and rescue operations in man overboard incidents. In: IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), pp. 57–64, UAE, Abu Dhabi, November 2020
11. Qingqing, L., et al.: Towards active vision with UAVs in marine search and rescue: analyzing human detection at variable altitudes. In: IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), pp. 65–70, UAE, Abu Dhabi, November 2020
12. Qiao, D., Liu, G., Lv, T., Li, W., Zhang, J.: Marine vision-based situational awareness using discriminative deep learning: a survey. J. Marine Sci. Eng. **9**(4), 397 (2021)
13. Bochkovskiy, A., Wang, C., Liao, H.M.: YOLOv4: optimal speed and accuracy of object detection. CoRR, vol. abs/2004.10934 (2020). https://arxiv.org/abs/2004.10934
14. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Computer society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 886–893 (2005)

15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**, 91–110 (2004)
16. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580–587 (2014)
17. Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision (ICCV) 2015, pp. 1440–1448 (2015)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28. Curran Associates Inc. (2015)
19. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788 (2016)
20. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
21. Redmon, J.: YOLOv3: an incremental improvement (2018)
22. COCO - common objects in context web site. https://cocodataset.org/. Accessed 20 June 2020
23. Tan, M., Pang, R., Le, Q.V.: EfficientDet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10 781–10 790 (2020)
24. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
25. Hraybi, S., Rizk, M.: Examining YOLO for real-time face-mask detection. In: Smart Cities Symposium (SCS 2021), vol. 2021. Institution of Engineering and Technology, pp. 571–575 (2021)
26. Katsamenis, I., Protopapadakis, E., Voulodimos, A., Dres, D., Drakoulis, D.: Man overboard event detection from RGB and thermal imagery: possibilities and limitations. In: Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments, ser. PETRA 2020. New York, NY, USA. Association for Computing Machinery (2020)
27. Redmon, J.: Darknet: Open source neural networks in C. https://pjreddie.com/darknet/. Accessed 14 Apr 2022
28. Padilla, R., et al.: A comparative analysis of object detection metrics with a companion open-source toolkit. Electronics **10**(3), 279 (2021)

Advancement in Manufacturing and Sustainability

Data Driven Decision Making When Transitioning Towards a Modular Setup

Morten Skogstad Nielsen, Thomas Ditlev Brunoe, Ann-Louise Andersen(✉), and Kjeld Nielsen

Department of Materials and Production, Aalborg University, Fibigerstraede 16, 9220 Aalborg, Denmark
ala@mp.aau.dk

Abstract. Modularization is becoming a more recognized strategy for manufacturing companies in order to stay competitive when customers are requiring more customized products at commodity prizes and with lead times as of-the-shelf products. When manufacturing companies decide to make a transition from an overall non-modular to a modular approach, one of the main considerations is the technical task of converting a portfolio of non-or-semi modular products and manufacturing systems into a number of modular architectures. Because products and manufacturing systems may have been developed over time and in different locations in the organization, information is rarely collected in one place and in a format which makes is suitable for analysis. Therefore, the purpose of this research is to present an overall approach for collecting, modelling and analyzing data in order to create an decision support system (DSS) to be used when answering some of the questions that will arise in the technical transformation process. This includes both the identification of data, the ETL process and the modelling of these data in order to perform advanced analytics. Finally some examples from a case company will be presented.

Keywords: Modularization · Data modelling · Decision making

1 Introduction

Modularization as a strategy or as a strategic initiative has become more frequent in manufacturing companies within the last decades [1–3]. These strategies and initiatives are driven by the ever increasing customer requirements for more customized products at the same price, speed and quality as mass produced products [4]. The task of introducing the concept of modularization in manufacturing companies are far from simple and affects both people, process, products and manufacturing across the entire value chain [5–7]. Manufacturing companies started with the concepts of platforms and modularity in the product domain more than 25 years ago in order to effectively develop and introduce a large number of product variants in short time, while reducing variety-induced complexity and gaining economies-of-scale [8–10]. When product platforms are developed and applied in practice, reuse of manufacturing equipment and process increases,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 597–606, 2023.
https://doi.org/10.1007/978-3-031-16281-7_56

thereby increasing the adjustability of the manufacturing system [14], but only more recently have the concept of platform-based manufacturing been introduced [11, 12]. Many of the manufacturing companies that embark on this modularization journey have not been able to reach a complete modular setup and thereby not been able to realize the expected benefits [13]. Hansen P. K. et. al. [13] concludes that a significant lack of methods and approaches within the modularization phenomenon causes these companies to fail and Brunch J. et. al. [14] identified information management as a critical challenge when integrating product and manufacturing system development. Several approaches for modelling product and manufacturing systems have been proposed. One approach is the configurable component method by Claesson [15] which provides a generic modelling building block for objected-oriented models in both the product and manufacturing domains. Another method proposed uses vertically aligned class diagram to combine product variant master with UML Class diagrams [16]. Lastly, an approach for integrated modelling of products and processes, enabling a link between a portfolio of product and components to a portfolio of manufacturing equipment and process, thereby enabling automated matching of components and manufacturing equipment is proposed [17]. Therefore the purpose of this research is to identify a method of how it is possible to create a system in which it is possible to collect data, create information, and conduct analysis in order to support decision makers. The focus in this case will be on how the manufacturing company can identify, collect and utilize its internal data from various IT systems to create both standard analysis that could be reused across multiple focus areas but also one-of-a-kind analysis for specific focus areas. The remainder of this is structured as follows: Sect. 2 describes the ontology, data model and the process. Section 3 presents some examples of both standard and one-of-a-kind analysis. Finally Sect. 4 presents the discussion, while Sect. 5 conclusively summarizes the research contribution and future research direction.

1.1 Methodology

To address the research objectives stated above, this research was conducted using a mixed method of design science and action research methodologies. Design science research was used because of its focus on both rigor and relevance while designing innovative artefacts to be used application domains e.g. organizations, people or systems [18, 19]. Action research was used because of its primary focus to produce practical knowledge that is useful to people in the everyday conduct of their life [20]. So to ensure relevance through action research, one of the researchers was placed in a manufacturing company for a period of 24 month. Through this collaboration the researcher was able to participate in development projects and gaining access to the full, raw and unfiltered company data from both ERP and other IT systems. By using action research and participating in multiple development projects within the company, the researcher was able to establish an in-depth knowledge within each project [21].

1.2 Case Company

The case company chosen for this research is a large manufacturing company, that operate on a global scale with approx. 19.000 employees and an annual turnover of approx. 3.5

B€. It has a long history of designing and developing both products and manufacturing systems internally and therefore holds allot of the insights and knowhow in both domains within the company. Previously this company have conducted modularization initiatives, however these initiatives has been initialized by engineers and product domain experts and only on subsections of the portfolio. There for, there are indications that there is knowledge an overall acceptance and willingness in the design and development part of the organization to embrace the concept of modularization. In the new 5 year company strategy modularization, was put on the agenda from top management and the first company vide modularization initiative was launched. This was done by creating a new department with the overall responsibility of creating and managing modular architectures in both the product and the manufacturing domain. The department works as the main driver for modularization and engages stakeholders across the entire value chain.

2 Ontology, Tables and Data Model

As the overall purpose with this data model is to support engineers and domain experts in creating modular architectures, the process of designing the model have been done through an iterative process with discussion between the domain experts and the model designer. This is done to ensure the model is designed with the capabilities required by the domain experts. Through the discussion with domain experts it has been clear that a holistic view of the business was necessary and therefore information from the three overall domains, market, product, and manufacturing was required. The proposed approach to model the overall business setup by using data describing market, product and manufacturing builds on a company-specific ontology based on how data are modelled in respective IT systems within the company. In Table 1 is a list of the finial tables as created by the ETL process as described by [22] and Fig. 1 display how the tables are modeled together in a UML Class diagram.

Table 1. List of tables created by the ETL process

Table name	Short description
Material master	List of unique materials in the company (both products and components, and internally produced and external sourced)
Material characteristics	Table displaying all characteristics for all materials in the company
Bill of material	Table creating the consist-of/ part-of relationship between materials creating BOM
Material class	Table used in company ERP to classify materials in a hierarchy
Cost	Table displaying all cost related to all materials in the company by date and location

(*continued*)

Table 1. (*continued*)

Table name	Short description
Production output	Table displaying all materials produced by location, date, volume, value
Scrap	Table containing scrap rate for all material in the production by date, volume, and value
Vendor	List of unique vendors
Purchasing	Table displaying all materials purchased by vendor, date, volume, and cost
Sales hierarchy	Table displaying hierarchy of all products by e.g., product family, product groups
Sales volume	Table displaying sale of all materials by, date, location, volume, and revenue
Warranty	Table displaying product groups returned from customer due to warranty by date, and location
Manufacturing footprint	Table displaying all production sites
Equipment master	Table displaying all unique equipment in the company
Equipment characteristics	Table displaying all characteristics for all equipment in the company
Equipment value	Table displaying how the equipment are physical located within production sites
Functional location	Table displaying how the equipment are physical located within production sites
Production resources and Tools	Table linking materials to equipment during production

2.1 Product Domain

In the product domain, the table 'Material Master' is used as a connection point for many of the other tables as each material is unique. Each material belongs to a specific material class in 'Material Class' table, and depending on which material class a material is assigned to a list of specific characteristics are assigned to each material in the 'Material Characteristics' table. Each material can then have multiple BOM´s depending on how many manufacturing sites are producing the given material, meaning one unique material can have multiple BOM´s containing different components. Each material can be produced in multiple manufacturing sites at different times and in different volumes which is in the 'Production Output' table. When materials are being produced scrap can occur in the production at different times which is in the 'Scrap' table. Each material will also have a cost which can be broken down into cost components (e.g. Material cost, Labor cost, Variable cost) and which is varying over time, manufacturing site and/or vendor. Each material can also be linked to multiple vendors if dual sourced. This is in the vendor table and when materials are purchased, date, amount, cost, vendor are registered in the 'purchasing' table.

2.2 Market Domain

In the market domain each material belongs to a material group which is the lowest level in the 'Sales Hierarchy' table. This is used for financial reporting based on sales as the sales hierarchy resembles the structure of the sales organization. For each material group it is registered how many products are returned with warranty issues and these are registered in the 'Warranty' table. Finally all sales volume, date and location is registered in the 'Sales Volume' table.

2.3 Manufacturing Domain

In the manufacturing domain the 'Equipment Master' table is used as connection point for all the tables as each equipment has a unique equipment-number. Here each equipment can have a list of characteristics similar to the 'Material Master' table. Each equipment will also be registered with an acquisition value and a date of acquisition which is registered in the 'Equipment Value' table. Next, each equipment belongs to a specific functional location depending on it´s physical location in the manufacturing site which is registered in the 'Functional Location' table. And each manufacturing site, which is registered in the 'Manufacturing Footprint' table, is divided in functional locations. Finally each equipment is linked to the 'Production Resources and Tools' table so that itis possible to see which equipment is used when manufacturing the materials in the 'Material Master' table.

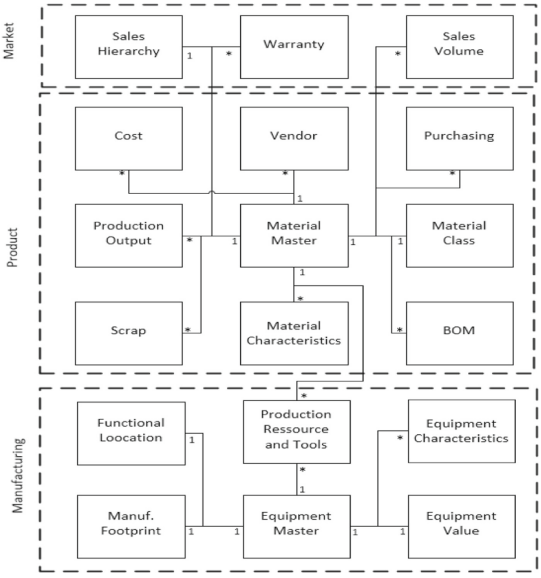


Fig. 1. UML class diagram describing the data model and the relationship between the different data tables

3 Examples

Through discussion with the different domain experts it became evident that the types of analysis needed could be divided into two types. There was the type of analysis that could be used across multiple domains which was mostly used in the preliminary investigations, and there was the deep technical analysis that was highly specialized only to be used one time within one domain. The following will give examples of both types of analysis and what could be concluded based on data the model.

3.1 Multipurpose Domain Analysis

As an initial analysis domain experts would often start by conducting a cost/performance analysis, looking at how a group of product performs on a set of performance parameters compared to its cost. Another analysis could be to create a heat map displaying the sales volume or the manufacturing cost for a specific group of product over a time period distributed over performance parameters. However in the cost/performance analysis, if the group of products found to be within scope was too big, the domain experts would typically select one or two specific product for each performance step and conduct the analysis on these as collecting data for all the products would be too time consuming. Meaning that the analysis would not be fully representative for the entire portfolio but only for those chosen by the domain experts. These types of analysis was previous done by the individual domain experts, meaning they would have to perform the entire ETL process from the company ERP system and typically make the analysis in Excel. This setup is far from ideal as it is both time consuming and the risk of making mistakes is high. Because this new data model is created in a cloud solution where the infrastructure is made so that all the data are weekly updated it was possible to use an 'of-the-shelf' BI tool to create a power script that combined the required data and made it possible for the domain experts to use 'drag-and-drop' to create these types of analysis always using the newest data. By using this BI tool the domain experts would not only save time by not having to collect data and combine it to perform the analysis, they would also get all the data available in the system and not risk missing out on data or combining it in a wrong way that could lead to a wrong decision being made. In Fig. 2 the process that domain experts now uses to create a heatmap is displayed.

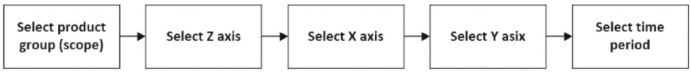


Fig. 2. The process domain experts uses to create heatmaps with the new data model

In a given example where the domain could be an electric motor and the domain expert would like to know what has been produced over the previous 12 month distributed on the performance parameters 'kilo watt' and 'motor efficiency'. The domain expert would simply choose from drop down menus as described in the top boxes in the flow diagram displayed in Fig. 2 and the system would collect data from the tables as described in the lower boxes in Fig. 3.

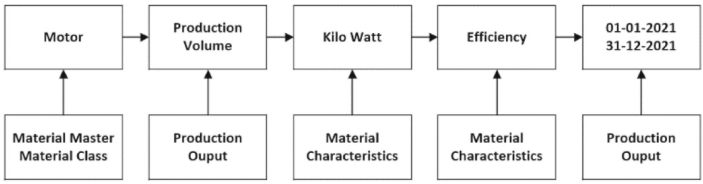


Fig. 3. Example of the process for an electric motor

These heatmaps could then be used to see if there are areas in the portfolio that are being produced in very low number and therefore should be excluded or covered by another architecture. If the sales volume or the revenue was chosen on the z-axis it would be possible to see if there are areas in the market which is not covered or if there are areas with is obsolete due to low revenue. Furthermore, if the profit was chosen on the z- axes it would be seen if there are areas in the portfolio which are not generating profit or even causing profit loses (Fig. 4).

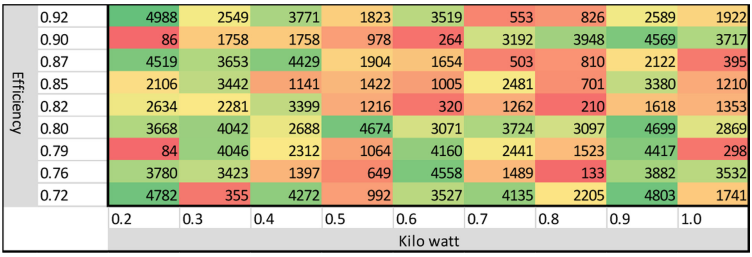


Fig. 4. Heat map displaying the production volume of motors in a given time period distributed on performance parameters 'kilo watt' and 'motor efficiency'

3.2 One-of-Analysis

The one-of-analysis are made based on specific requests from domain experts. These often came as a highly technical questions and then needed to be broken down in order to identify which data to use, how to combine them, and what analysis to perform. In the following example the questions from the domain expert was: for a group of products that contain several product families, the component 'shaft' defines a key interface, how many unique interfaces is there across these product families? From a discussion with the domain expert it was learned that for a shaft it is the shaft type (Spline or normal) and the outer diameter that defines the interface to the remaining system. Based on these information it was possible to create a process that could answer the question. In Fig. 5 is the process that was made and from which tables data was used in order to answer the question.

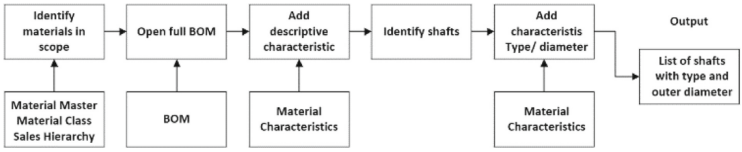


Fig. 5. The process created to identify the number of unique interfaces from shaft to remaining system.

The first step in the process was to identify all the materials in scope. This was done through the material classes and the material groups from the 'sales hierarchy' table. Once all the materials was found, the second step in the process was to open up all the BOM´s down to the lowest level of each BOM (in this case hundreds of BOM´s was opened). The third step of the process was to join a specific characteristic from the material characteristics table on the components in the BOM. This characteristic is a description/categorization that is made on all materials when they are created in the ERP system based on a certain ruleset. The fourth step in the process was to use this descriptive characteristic to search out the finished shaft in the BOM. Now all the unique shafts that was used across these product families was identified. The last step was to join the two characteristics, type and outer diameter to the list of unique shafts and then it was possible to group them first by type and by outer diameter. Here it became evident that there was a high number of different outer diameter sizes, causing a high level of unnecessary complexity in the remaining architecture. Once the variance in the shaft type and outer diameter was found the domain expert would have to identify which of these that where to be used going forward and decide whether the existing product that was not using the new shaft standard should be redesigned in order to reduce complexity.

4 Discussion

The applicability of the method presented in this research is highly independent on the company context in which it is applied. The data model and the ontology on the other hand is highly dependent on both the company and the context in which it is applied. That is because this both depend on how the existing data and IT systems are structed in its current form within the company and for what purpose the model is designed. Although this paper presents only one case company, it serves as an illustration of how this method could potentially be used in other industrial settings. Nevertheless, having only researched one case a few limitations can be identified. Because of the complexity in today´s IT infrastructure within companies and the high level of skill required to create a model like the one presented in this research the collaboration between the data scientist and the domain experts require attention in order to translate highly technical questions into something quantifiable within the data set. Another limitation would be the resources that will need to go into creating and maintaining the infrastructure required to use this method. In this case this was found to be extremely time consuming and required a broad range of stakeholders as the model contains data from various parts of the organization thereby involving a high number of data owners. However, when the infrastructure and the data model are created and in operation the benefit for the company are extremely

high. This is because this data repository and data model become one point of access and one point of truth that works across departments so that when engineers and domain experts are to perform an analysis there is one please where they can turn to and acquire the data they need in a fast and reliable way.

5 Conclusion

In this research, a new approach was presented where the collection of data from multiple systems enabled the creation of a data model that assisted engineers and domain experts in the modularization process. Implementing and maintaining a data model like the one presented in this research can have major implications on a company as it would aid the use of data analysis when taking decisions. Through the case it was shown that creating a model like presented in this research had the ability to assist in both 'business' decisions but also in answering highly 'technical' questions. Future research extending on what is presented in this paper could focus on several topics. Firstly, having a data model like the one presented available should enable more advanced analysis, so one research direction could be to identify what types of analysis could be created and how could these help business become more profitable. Secondly, what effect, if any would the easily availability of valid data have on NPD and other critical business process

References

1. Løkkegaard, M., Mortensen, N.H.: Assessing the financial potential for modularization: a case study in a global OEM. In: Proceedings of the International Conference on Engineering Design, ICED, vol. 3, pp. 21–30 (2017)
2. Thyssen, J., Israelsen, P., Jørgensen, B.: Activity-based costing as a method for assessing the economics of modularization-a case study and beyond. Int. J. Product. Econ. **103**, 252–270 (2006). https://doi.org/10.1016/j.ijpe.2005.07.004
3. Sanchez, R.O.N.: Competition. **16**, 135–159 (1995)
4. ElMaraghy, H., Schuh, G., Elmaraghy, W., et al.: Product variety management. CIRP Ann. Manuf. Technol. **62**, 629–652 (2013). https://doi.org/10.1016/j.cirp.2013.05.007
5. Hansen, P.K., Sun, H.: An incremental approach to support realization of modularization benefits. In: IEEM2010 - IEEE International Conference on Indus-trial Engineering and Engineering Management, pp. 173–177 (2010) . https://doi.org/10.1109/IEEM.2010.5674448
6. Östgren, B.: Modularisation of the products gives effect in the entire production (1994)
7. Sanchez, R., Mahoney, J.T.: Modularity, flexibility, and knowledge management in product and organization design. Strateg. Manag. J. **17**, 63–76 (1996). https://doi.org/10.1002/smj.4250171107
8. Erens, F., Verhulst, K.: Architectures for product families. Comput. Ind. **33**, 165–178 (1997). https://doi.org/10.1016/s0166-3615(97)00022-5
9. Ericsson, A., Erixon, G.: Controlling Design Variants: Modular Product Platforms. Society of Manufactuting Engineers (1999)
10. Simpson, T.W., Siddique, Z., Jiao, J.R.: Platform-based product family product platform and product family design. Meth. Appl. 1–15 (2005)
11. Michaelis, M.T., Johannesson, H.: From dedicated to platform-based co-development of products and manufacturing systems. Enab. Manufact. Competitiveness Econ. Sustainab. 196–202 (2012) . https://doi.org/10.1007/978-3-642-23860-4_32

12. Michaelis, M.T., Johannesson, H.: Platform approaches in manufacturing - considering integration with product platforms. In: Proceedings of the ASME De-sign Engineering Technical Conference, vol. 9, pp. 1115–1124 (2011). https://doi.org/10.1115/DETC2011-48275
13. Hansen, P.K., Sun, H.: A comprehensive view on benefits from product modularization. In: 11th International Design Conference, DESIGN 2010, pp. 233–242 (2010)
14. Bruch, J., Bellgran, M.: Integrated portfolio planning of products and production systems. J. Manufact. Technol. Manag. **25**,155–174 (2014) . https://doi.org/10.1108/JMTM-09-2013-0126
15. Claesson, A.: A configurable component framework supporting platform-based product development. Doktorsavhandlingar vid Chalmers Tekniska Hogskola 1–159 (2006)
16. Haug, A., Hvam, L., Henrik, N.: Computers in industry a layout technique for class diagrams to be used in product configuration projects. Comput. Ind. **61**, 409–418 (2010). https://doi.org/10.1016/j.compind.2009.10.002
17. Brunoe, T.D., Andersen, A., Sorensen, D.G.H., et al.: Integrated product-process modelling for platform-based co-development. Int. J. Product. Res. **58**(20), 6185 6201 (2020). https://doi.org/10.1080/00207543.2019.1671628
18. Hevner, A.R.: A three cycle view of design science research. Scand. J. Inf. Syst. **19**, 87–92 (2007). http://www.aisel.aisnet.org/sjis/vol19/iss2/4
19. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information research 1. Des. Sci. IS Res. MIS Q. **28**, 75 (2004)
20. Kagan, C., Burton, M., Siddiquee, A.: The handbook of action research - introduction. In: Handbook of Qualitative Research Methods in Psychology, vol. 468 (2006)
21. Leonard-barton, D.: A dual methodology for case studies: synergistic use of a longitudinal single site with replicated multiple sites author (s): dorothy leonard-barton source : organization science, 1, 3. Spec. Issue Longitudinal Field Res. M. **1**, 248–266 (1990)
22. Dayal, U., Wilkinson, K.: Data integration flows for business intelligence. In: Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, pp. 1–11 (2009)

A Novel Method for Component Positioning in Thermoformed Electronics

Behnam Madadnia(✉), Frederick Bossuyt, and Jan Vanfleteren

Centre for Microsystems Technology, imec and Ghent University, Technology Park 126, Gent-Zwijnaard 9052, China
{behnam.madadnia,jan.vanfleteren}@ugent.be,
Frederick.Bossuyt@UGent.be

Abstract. Curve-shaped electronics have been on the rise in recent years. One crucial challenge in this area is reproducibility, which means in the different fabricated samples, the position of the electronic components should be the same. This paper aims to present a reliable approach for accurate component positioning in thermoformed electronics. We have proposed a non-stretchable conductive structure for accurate positioning of the electronic components, which can assure the conductive routes' mechanical stability during lamination and thermoforming of the substrate. We have verified our approach's accuracy by applying our method on a real industrial luminaire mold for positioning six LEDs in the center of the luminaire parts. We have measured the LEDs' position after thermoforming to prove the design's repeatability. The experiment results state that the proposed method is capable of positioning electronic components in thermoformed 3D electronics and 400% improvement is reported in comparison with the previous fabrication approaches.

Keywords: Thermoforming · 3D electronics · Positioning

1 Introduction

For several years, 3D shaped electronics have been on the rise, with many uses in home appliances, the automotive sector, and manufacturing. Curve-shaped electronics can be manufactured using a variety of technologies, such as In-mold electronic technology (IME) [1, 2], Stretchable mold interconnect (SMI) [3–7], 3D-MID [8]. IME is a method that integrates traditional plastic molding with conductive flexible or even elastic inks printed on plastic substrates. SMI technology uses stretchable metal conductors and thermoforming to realize the 3D structure, which is the main focus of this paper. In 3D-MID approach, standard processing steps, including laser structuring and copper plating, are used. Vanfleteren et al. investigated and compared the three methodologies [9]. There is no control over the ending position of the components during thermoforming in SMI and IME technology. Considering that thermoforming applies a complex deformation to the 2D substrates as a result. Consequently in most applications, implementing a repeatable

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 607–615, 2023.
https://doi.org/10.1007/978-3-031-16281-7_57

procedure is impossible. The biggest obstacle of 3D-shaped electronics, which makes it challenging for industrialization, is the repeatability of the electronic components' spatial position. The critical point in the component positioning in such thermoformed pieces is to determine where to position the element in the 2D design so that after thermoforming the entire flat surface, the components end up in the desired position in the 3D sample. For example, in Fig. 1, the experiment aimed to position the circled-islands in each section's center. It can be seen that the islands' spatial locations are at various points after thermoforming, meaning that precise positioning can't be controlled by using the meander shaped stretchable conductors of this design.

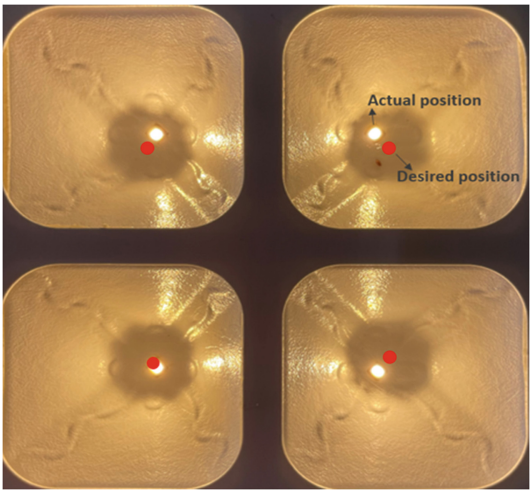


Fig. 1. Thermoformed sample for luminaire mold, red dot represents the desired position and the light spot shows the actual LED position.

There is a limited amount of research in the field of electronic component positioning in thermoformed and 3D shaped electronics. A patent application by Vanfleteren has used the "Degrees of freedom" methodology to position the electronic components in thermoformable electronics, by using this approach the meanders will guide the component in just one possible spatial position [10]. Jonathan Ting recently demonstrated software that allows 2D-to-3D mapping, guided structural and electrical circuit design, and an interactive user interface in partnership with Microsoft [11]. To transfer 2D produced circuits to 3D shapes, some mathematical mapping was used in the paper. By altering the component position in 2D, one may see the new position in 3D interactively in the software. The paper has not considered the physics of the thermoforming and just investigates mathematically the subject. Besides, this approach can only be used for IME technology, not for SMI because the printed electronic tracks deform in the same way as the substrate they are printed on but for SMI, the stretchable interconnects deform differently compared to the substrate. Jungrak Choi, recently, represented a customizable, conformable, and stretchable fabrication method of 3DE based on predistorted

pattern generation and thermoforming. Through, this method, custom-designed 3DE is fabricated through the thermoforming process. The fabricated 3DE has high 3D conformability because the thermoforming process enables the complete replication of both the overall shape and the surface texture of the 3D mold. Furthermore, the usage of thermoplastic elastomer and a liquid metal–based conductive electrode allows for high thermoformability during the device fabrication as well as high stretchability during the device operation [12].To the best of our knowledge, there is no study of component positioning in SMI technology.

In this paper, we propose an approach for positioning the electronic components in structured flexible circuit technology without stretchable parts. We have used a non-stretchable structure as the conductor in the design. These structures remain mechanically stable during the thermoforming, which means non-stretchable parts will have the same length before and after forming. Instead of stretching, these non-stretchable parts slide and bend through the polymer. Using this approach, we eliminated all kinds of uncertainty in the copper pattern's length, which can be caused during thermoforming. This idea is submitted as a European patent application. First, we have made a simple experiment to prove the concept of using non-stretchability in thermoformed electronics. In the next part, we've investigated the fabrication process and thermoforming of the 3D shaped circuit. Finally, we have assessed the accuracy and repeatability of the fabrication process. Experimental results show that this approach is a repeatable and high-precision method for component positioning in thermoformed electronics.

2 Proof of Concept

We introduce a design method, which assures the repeatability of electronic component's position in thermoformed samples with particular forms. The critical point in the component positioning in such thermoformed pieces is to determine where to position the element in the 2D design of the flat surface so that after thermoforming, the components end up in the desired position.

2.1 Design Process

We want to place a LED in the center of a luminaire mold. Using SolidWorks software, we have designed the luminaire mold in Fig. 2 (A, B). The goal is to develop a copper pattern in 2D, which can guide the LED position into the center after thermoforming. By measuring the $L1 = 96$ mm (polymer's final length) and $L2 = 50$ mm (polymer's initial length) in Fig. 2.C, it is clear that there is a specific elongation, which is the proportion of polymer's surface length before and after thermoforming, based on the mold after thermoforming. In Fig. 2.C, the polymer's surface length before and after thermoforming can be seen, and elongation is 92% ($\frac{L1-l2}{L2}$). As mentioned in the previous solution for positioning [10], it may seem that because the polymer stretches during thermoforming, the conductive structures should also be stretchable; otherwise, the conductor may break. In [10, 13], the meander shape design is considered as the conductive track because of the

stretchability. Considering that this kind of design causes lots of positioning uncertainty after thermoforming, and the length of the structures randomly vary due to the lab thermoforming process such as limited uniformity in heating, the LED position can't be in the same place for each sample. As a result, we will not have a repeatable result. We present a non-stretchable structure as the conductive track, which slides in between polymer layers instead of stretching. Non-stretchability is the crucial feature because it can ensure the same length and mechanical stability of the conductive structure after forming. The distance between the desired position and the mold's outer edge can be easily calculated, "Distance" = 48 mm, as shown in Fig. 2.D.

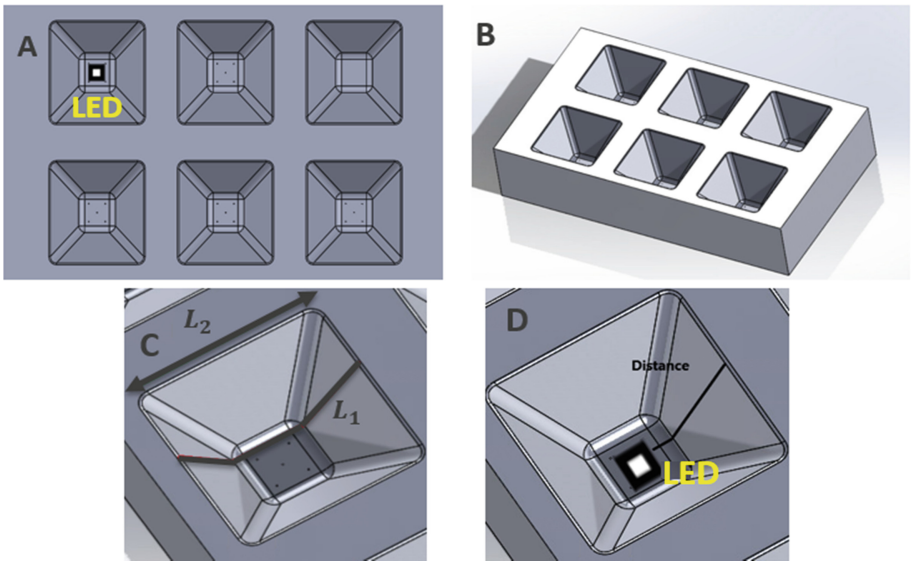


Fig. 2. (A) Top view of the luminaire mold, (B) Perspective view of the luminaire mold, (C) ***L***1, the polymer's surface length after thermoforming and ***L***2, before thermoforming, (D) Distance between the desired position and the mold's outer edge (Distance = 48 mm).

2.2 Design Process

We have designed a non-stretchable structure using AutoCAD with a length of 48 mm, as shown in Fig. 3.A. After laser cutting the pattern on Upisel SR1220 (50 μm PI - 18 μm RA Cu), we laminated the lasered-copper design with the symmetric stack as shown in Fig. 3 (B, C, D).

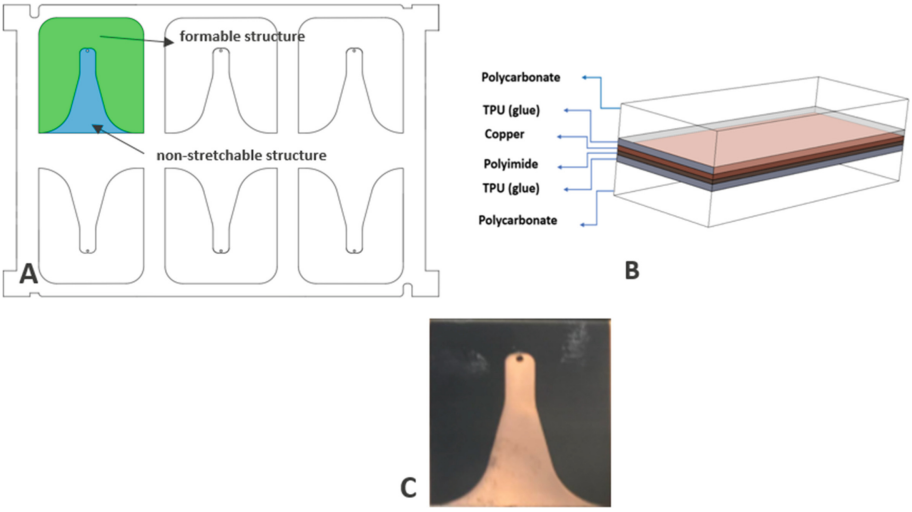


Fig. 3. (A) CAD design of the Non-stretchable structure with a length of 48 mm, (B) Symmetric lamination stack- perspective view, (D) Symmetrically laminated stack before thermoforming.

2.3 Thermoforming

Figure 4 shows the thermoforming process: it consists of aligning a prefabricated stack in the vacuum forming machine and heating the polymer to reach its glass transition temperature (A, B). The material will be sucked into the mold cavity by a vacuum as the room-temperature mold is lifted towards the polymer. When the polymer comes into contact with the cold mold, it begins to solidify, and the polymer becomes solid (C).

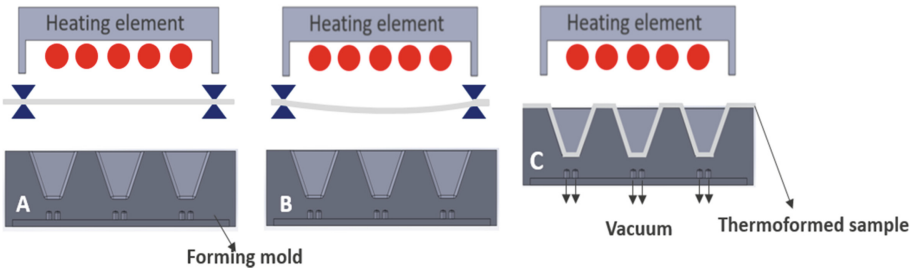


Fig. 4. Thermoforming machine heats a laminated stack until it reaches the thermoplastic polymer's glass transition temperature and becomes soft. Then, utilizing a strong vacuum, it is formed over a forming tool.

Figure 5 represents the final thermoformed sample. As a qualitative assessment, it is clear that the hole, mimicking a component, is exactly in the center of the pyramid. In the next part, we did a quantitative assessment of the thermoformed samples.

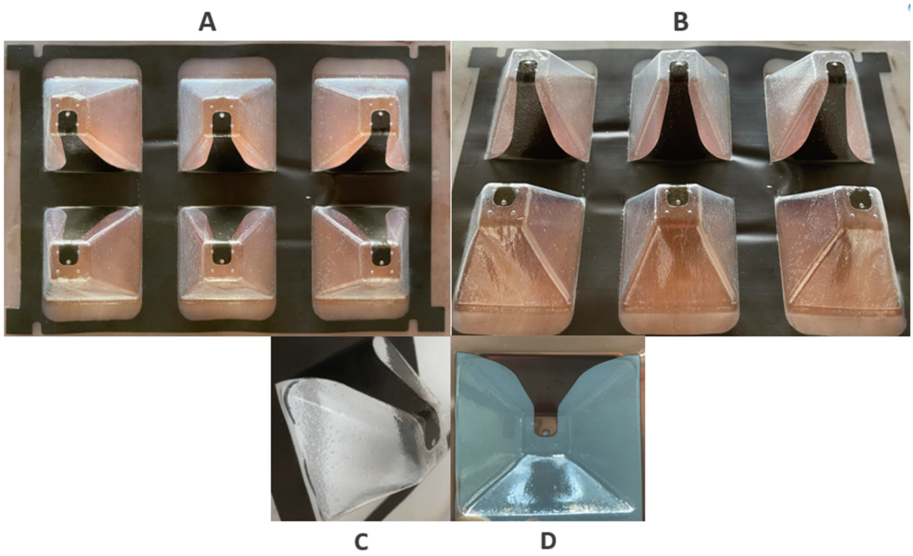


Fig. 5. (A) Backside view of the thermoformed sample, (B) Perspective view of the thermoformed sample, (C, D) Close up view of the strip in the formed sample.

2.4 Accuracy Analysis

We have used an optical microscope to have a quantitative assessment of the accuracy of the positioning method, which is fabricated and thermoformed in the previous section (Fig. 5). We measured the error in the X and Y direction using the microscope software (Fig. 6). Based on the measured data for six different luminaire parts, as an average error,

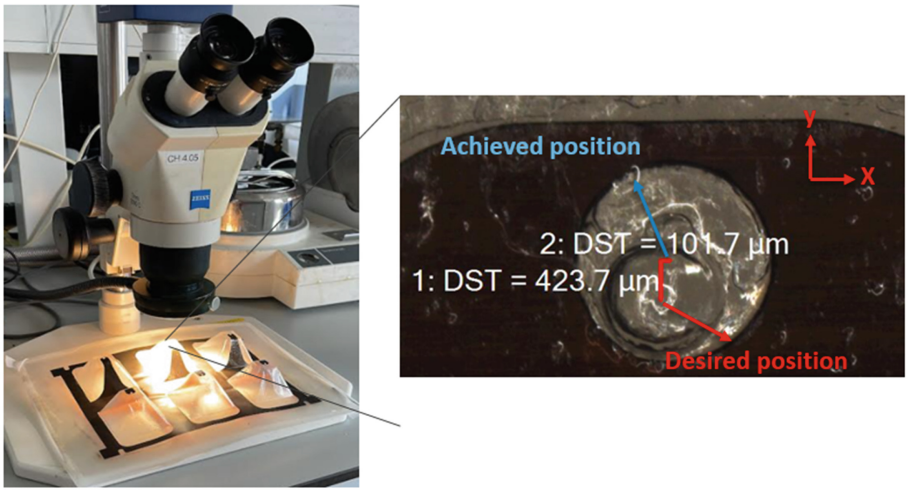


Fig. 6. Measurement setup: "Achieved position" are the coordinates of the center of a small circle in the flexible substrate; "desired position" are the coordinates of a small circle in the thermoformed polymer, implemented in the mold.

we can report 479 μm in the Y direction and 81.5 μm in the X direction (Table 1). We have also measured the position of the island in the fabricated sample using stretchable structures (Fig. 1) (Table 1).

Table 1. Measurement results for non-stretchable and stretchable approach

	X-Error (μm) Non-stretchable	Y-Error (μm) Non-stretchable	X-Error (μm) stretchable	Y-Error (μm) stretchable
1	186	932	423	1644
2	84	745	1831	508
3	0	440	3729	1102
4	101	118	2034	1390
5	0	305	1695	2763
6	101	423	3119	1797
Average	81.5	479	2138.5	1534

2.5 Improvement with Respect to the Existing Literature

We have compared the measurement results of the two different approaches by using the following criterion (formula 1) as depicted in Fig. 7.A. We have used this criterion to compare the accuracy of the two different approaches with each other. Figure 7.B shows a significant improvement in the accuracy of the positioning (e.g., 442% improvement in the positioning accuracy).

$$\text{Distance error} = \sqrt{X_{error}^2 + Y_{error}^2} \tag{1}$$

Using (1), we have calculated distance error for Non-stretchable and stretchable approach and respectively the values are 486.7 μm and 2631.7 μm. We can see about 400% improvement in the accuracy of the positioning Fig. 7.

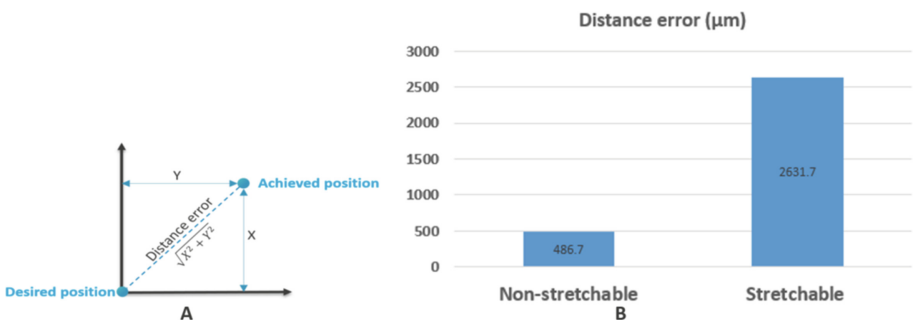


Fig. 7. (A) Represents distance error calculation between desired and achieved position, (B) The comparison between the accuracy of the non-stretchable and stretchable approach.

2.6 Elongation

There is a criterion called elongation in thermoformed parts, which is the proportion of the initial formable part to the final formed area. We should maintain a low value for elongation to secure the quality of the sample and the thickness distribution across the sample. We introduce the formula (2) for calculating the elongation. The different areas are depicted in Fig. 8.

$$\begin{aligned} Elongation &= \frac{((Area_2 - Area_3) - Area_1)}{Area_1} \\ &= \frac{((4234.47 - 810.25) - 1684.5)\text{mm}^2}{1684.5\,\text{mm}^2} = 199\% \end{aligned} \tag{2}$$

This means that for this design the area of freestanding polymer (not supported by the flexible polyimide substrate) has almost doubled due to the thermoforming step. Consequently the average polymer thickness has decreased with a factor of 2. We need to find the best possible width for the non-stretchable structure. Based on elongation, the larger the non-stretchable structure, the larger elongation; on the other hand, to maintain the mechanical stability of the non-stretchable part, the bigger width, the better. In conclusion, we should make a trade off between the mechanical stability (e.g., positioning accuracy) and the elongation.

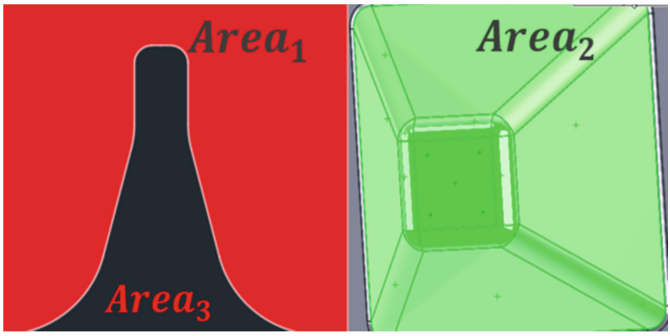


Fig. 8. Symmetrically laminated stack after thermoforming.

3 Conclusion

In conclusion, a positioning method to accurately and repetitively position electronic components in thermoformed electronics was proposed. We have used non-stretchable structures to establish our positioning method. We design the 2D stack so that there is just one possible 3D position for the component after forming. Measurements done by an optical microscope shows there is a maximum distance error of 486.7 μm, which is 400% improvement in comparison with the previous approaches. We have shown the accuracy of the presented method has significantly improved the results in comparison with the existing literature. Future studies can focus on the optimized width of the non-stretchable part.

References

1. Merilampi, S., Björninen, T., Haukka, V., Ruuskanen, P., Ukkonen, L., Sydanheimo, L.:¨ Microelectron. Reliab. **2010**, 50 (2001)
2. DuraTech. In-Mold Electronics (IME) (2019). https://www.duratech.com/capabilities/ime-in-mold-electronics. Accessed 30 Nov 2019
3. Chtioui, I.: Arbitrarily shaped rigid and smart objects using stretchable interconnections. IEEE Trans. Compon. Packag. Manuf. Technol. **6**(4), 533-544 (2016)
4. Plovie, B.: Arbitrarily Shaped 2.5D circuits using stretchable interconnects embedded in thermoplastic polymers. Adv. Eng. Mater. **19**(8), 1700032 (2017)
5. Vanfleteren, J., et al.: Free-form 2.5D thermoplastic circuits using one-time stretchable interconnections. Mater. Res. Soc. Symp. Proc. **1798** © 2015 Materials Research Society (2015)
6. Plovie, B., Vanfleteren, J., Vervust, T., Quintero, A.V., Bossuyt, F.: Design automation of meandered interconnects for stretchable circuits. IEEE Trans. Comput.-Aid. Des. Integr. Circ. Syst. **38**(9), 1648-1660 (2018)
7. Plovie, B.: Thermoplastic Electronic Circuits: Design, Technology and Characterization, Bart Plovie, Ghent University (2019)
8. Franke, J.: Three-Dimensional Molded Interconnected Devices (3DMID), 335 pages. Carls Hanser Verlag, Muenchen (2014)
9. Vanfleteren, J., Bossu yt, F., Plovie, B.: Flex PCB based technology for randomly shaped circuits. In: 2017 EIPC conference Proceedings (2017)
10. https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2018054623
11. Ting, J., Zhang, Y., Yoon, S.H., Holbery, J.D., Ma, S.: iMold: enabling interactive design optimization for in-mold electronics. In: CHI 2020, pp. 25–30, Honolulu, HI, USA (2020)
12. Choi, J., et al.: Customizable, conformal, and stretchable 3D electronics via predistorted pattern generation and thermoforming. Sci. Adv. **7**(42), eabj0694 (2021)
13. Vanfleteren, J., Bossuyt, F., Plovie, B.: A new technology for rigid 3D free-form electronics based on the thermoplastic deformation of flat standard PCB type circuits. In: 12th International Congress Molded Interconnect Devices (MID), pp. 1–4 (2016)

Towards Circular Production Systems: Outlining the Concept, Challenges and Future Research Directions

Filip Skärin(✉), Carin Rösiö, and Ann-Louise Andersen

Jönköping University, Gjuterigatan 5, 553 18 Jönköping, Sweden
filip.skarin@ju.se

Abstract. Manufacturing companies must significantly reduce their negative environmental impact, while simultaneously needing to be capable of managing changing customer requirements and increasingly fluctuating volume demands. These constraints require discontinuing developing production systems in line with the outdated and unsustainable linear economy, wherein the system is design and developed for short term production only. In this paper, it is motivated how a production system is distinguished from a product and why state-of-the-art circular economy theory must be adapted to production systems in order to deal with these issues. Furthermore, based on theory and inputs from workshops with participants from the industry, a conceptual framework for circular production system is proposed. In the proposed circular production systems, the principles of circular economy is adapted in order to maximize resource utilization and prolong the lifetime of the production system. Challenges towards the circular production systems have been identified through workshops and potential solutions described. Lastly, future research directions for further developing the circular production system is presented.

Keywords: Circular production system · Circular economy · Circularity · Reconfigurability · Production

1 Introduction

The linear economy has dominated largely throughout the past 150 years and is characterised by a single-loop product and material view where resources are extracted from earth, processed into a product, used and finally disposed [1, 2]. Nevertheless, due to the accelerating global warming crisis, the linear economy has started to be replaced by the concept of circular economy (CE). In a CE, the aim is to maximise the utilization of resources and prolong the lifetime of resources, in order to achieve a closed material loop [3–5]. Although many actors are vital in a CE, it is critical for manufacturing companies to convert the business into a CE, as they are responsible for a significant part in the global warming crisis. As an example, manufacturing companies in the United States were responsible for 23% of the country's total CO_2-e release into the atmosphere in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 616–625, 2023.
https://doi.org/10.1007/978-3-031-16281-7_58

2019 [6]. Even though manufacturing companies are in need of reducing CO_2 emissions, additional factors such as customer demands, technological advancements and global competition are influencing manufacturing companies' survivability and competitiveness [7]. Thus, it is critical that manufacturing companies are able to develop and manage production systems that are both sustainable, supports CE, as well as capable of being dynamically changed towards customer and market demands. However, while the state-of-the-art theories within CE, such as the 10R framework by Potting et al. [8], have an established product focus, the production system has yet to be in centre of attention. Producing the products in accordance to sustainable development, i.e. with *"…reduced negative environmental impacts, offer improved energy and resource efficiency, generate minimum quantity of waste, provide operational personnel health while maintaining and/or improving the product and process quality with the overall life-cycle cost benefits*" [9, p. 10], might play an important part. However, as CE is based on prolonging the lifetime of resources taken from earth, the production system needs to be regarded as well, not solely the products produced. Hence, studying production systems from the same lens of circularity should be considered as a key to achieving circularity. Therefore, the purpose of this paper is to propose how CE concepts can be applied in the production domain in order to bridge the challenges and to facilitate the understanding of circularity to the domain of production systems. This will incorporate aiding the transition of industry, identify challenges, make an outline of the concept, and suggest future research directions.

2 Theoretical Background

2.1 Circular Economy

In line with the increasing global warming crisis, the concept of CE has gained a significant increase in attention [3–5]. For instance, in academia, the number of publications published per year within CE increased more than 2.5 times between 2010–2015 [10]. However, the definition of CE has been debated. In attempt to conceptualizing CE, Kirchherr et al. [11] conducted an analysis of 114 definitions of CE. Their compounded definition involves CE described as an "*economical system, which "replaces the 'end-of-life' concept with reducing, alternatively reusing, recycling and recovering materials in production/distribution and consumption processes*" [11, p. 229]. The general aim with the CE is found to be connected to achieving sustainable development by prolonging the lifetimes of resources and, thus, minimizing the usage of raw materials extracted from the planet [11]. Related to this aim, CE has also been described as: "*a regenerative system in which resource input and waste, emission, and energy leakage are minimised by slowing, closing, and narrowing material and energy loops"* [12, p. 759].

Realizing CE is related to activities such as reduce, reuse, remanufacture, and recycle, frequently described as the R-methodology. The R-methodology stemmed from lean manufacturing and reducing waste [13]. During the 1990s, an evolved R-methodology including reduce, reuse and recycle (3R methodology) gained attention due to companies attempting to realize a greener manufacturing. In line with the accelerating global warming crisis, the 3Rs are nowadays reckoned as being insufficient for ensuring sustainable manufacturing [14, 15]. Instead, it has been argued that further Rs need to be

added. Initially the 6R framework was proposed, which includes the addition of redesign, remanufacturing and recover. The 6Rs have been found as an enabler for closing the resource and material loop within supply chains and, thus, creating the foundations for sustainable manufacturing [15]. However, the development of the R-methodology has continued and a framework consisting of a total of 10Rs [8] has been proposed, whereas refuse, rethink, repurpose and recover has been added (Fig. 1). In the extensive 10R framework, the Rs are divided into three main categories: 1) efficiently use of products and manufacturing operations, 2) product lifecycle extension to fit several lifecycles, and 3) maximise the material usefulness [8].

Fig. 1. 10R framework, adapted from Potting et al. [8].

However, throughout the evolution of the R methodology, various terms for describing the Rs have been used. In the 10R framework, Potting et al. [8] use similar terms as in the 6R framework, with some alterations. This implies that the R methodology is continuously developing and adaptable. Nevertheless, even though the 10R framework focuses on products, and the only R directly related to production is Reduce, it might be relevant to regard production systems through a lens of circularity. Thus, the need to compare and differentiate circularity from the traditional product perspective to the production system domain could be realized.

2.2 Comparing Product and Production System

It is motivated to adopt a circularity perspective also related to the production systems and not only to the product to be produced. However, there are major differences between a product and a production system, and the circularity concepts needs to be adapted to enable production system circularity. A product includes all systems and physical objects, including all associated services, that are produced or refined by humans. A production system can be described as all activities and resources needed to transform raw material into products. One might argue that a production system also can be regarded as a product, i.e., an artifact that is developed and realized. However, a production system is complex comprising not only physical hardware, but also people managing and operating the hardware and responsible for information sharing within the production system [16]. The production system includes several fundamental parts that all must act together to achieve the purpose of the system [17]. The production system has been described as the arrangement and operation of machines, tools, materials, people, and information that are required to manufacture a product or part of a product [16, 18]. In order to change the system and to prolong its lifetime an ability to restructure the production system through rearrangement and reuse is needed [19, 20]. Thus, to enable circularity in a production system implies other abilities/characteristics than for a product.

While there is increased focus on the need for circularity, the need for designing a production system ready to be changed has for a long time been regarded as crucial [21, 22]. Within the field of changeable and reconfigurable production principles for prolonging the lifetime of the production system through the enablers of modularity, integrability, diagnosability, convertability, scalability and customization are described [7]. This provides abilities in the system for restructure and reuse which has been motivated by sustainability improvements e.g. [21, 22].

The term circular production system used in this research should not be confused with the term circular manufacturing system, which for instance is described by Asif [23] and Lieder [24]. In their research, a circular manufacturing system is defined as a system which is designed with the purpose of enabling the possibility to manage a product throughout several lifecycles [23]. This definition implies a product focus, as in the 10R framework, whereas the manufacturing system's purpose is to enable reusability of the product.

To conclude, the aim with the circular production system is to maximize the utilization of resources used in the system, as well as the resources which the system itself consist of. Thereby, the Rs are in need for adaptation to the production system domain, as a means for developing these circular production systems (see Fig. 2).

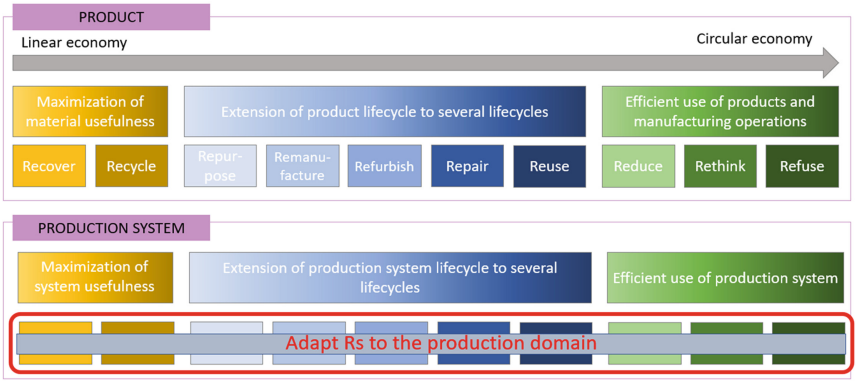


Fig. 2. Applying the 10R framework to the production domain, based on Potting et al. [8].

3 Method

The conceptual framework proposed in Sect. 5 was developed through an iterative process, whereas an initial version of the framework was developed based on theory (10R framework). Two workshops were carried out in order to gain insights into how company representatives view and practice circularity in production, thus, aiding in conceptualizing the framework. The aim with the workshops was also to identify future challenges in developing circular production systems and gain practically derived insights of the framework. In between the workshops the framework was continuously developed, however, after the second workshop, the conceptual framework was adjusted in order to cope with some of the identified challenges, resulting in the version presented herein. The workshop was chosen as a data collection technique due to the possibility to gain new insights and ideas through the discussions between participants [25], which was deemed necessary in order to adapt CE to a production domain and to identify challenges in the conceptual framework. The workshops were carried out online with key representatives with positions specialized in production systems and production development from 6 companies and organizations. In both workshops, two of the authors were participating, one assigned a facilitator role, the other taking notes [25]. The first workshop consisted of 12 participants and was carried out in October 2021 with a focus on sustainable production. The workshop covered the main principles for establishing sustainability in production as well as the foundations for CE. The workshop consisted both of theoretical introductions to the aforementioned topics and discussions within the areas. The discussion questions were predetermined and participants were divided into smaller groups to discuss the questions. The insights from the discussions were thereafter shared amongst all participants in the workshop and a joint discussion facilitated by the authors occurred. The second workshop consisted of 7 participants and was carried out in March 2022. This workshop functioned as a continuation of the previous workshop, and covered the topic circular production development. The workshop was carried out with the same structure as the first workshop.

4 Challenges Toward Circular Production Systems

Several challenges towards circular production systems have been identified:

- Lacking competence in purchasing
- Lacking knowledge of how to design and develop circular production systems
- Inability to reconfigure and continuously update the production system
- Lacking collaboration between manufacturing company and system supplier
- Uncertain decision making in prolonging the system's lifetime
- Future legal demands

One of the challenges towards circular production systems include the ***lack of competence amongst purchasing personnel***. The participants stressed the need for purchasers to broaden the scope when purchasing from only including current needs, i.e. purchasing production systems which solely are capable of producing current products, to also include added changeability in the production system to also be able to manage future products as well. Hence, adapting outdated and linear economy descending purchasing models to this novel view might be required in order to enable the design and development of circular production system.

Challenges regarding an ***existing lack of knowledge of how to design and develop circular production systems*** and current ***inability to reconfigure and continuously update/upgrade the production system*** were also mentioned. Many current production systems are found incapable of supporting circularity, as they are dedicated and, thus, do not possess the ability for a continuous update. This might be considered a key solution in circular production systems, as changeability in the production systems, whereas solely a part of the production system is required to be altered in order to adapt the production system into being capable of managing an expected need for managing more frequent product introduction and fluctuating demands. Hence, a rather significant change in how current production systems are designed and continuously developed needs to be realized. The proposed conceptual framework might be recognized as an instigation in this change.

Furthermore, establishing a ***long-term collaboration between manufacturing company and production system supplier*** was also identified as a challenge towards circular production systems. In order to facilitate a continuous update of the production system, and, thus, avoid investments in new production system when e.g. introducing new products, a long-term collaboration between these actors might be required.

Further, also the challenge of ***uncertain decision making*** was mentioned, for instance in terms of the decision whether to purchase new, renovate or rebuild the production system when breaking down. The companies experienced a lack of knowledge regarding these types of decisions. Creating decision-supporting models and tools, e.g. for comparing and calculating the financial result and environmental impact for the options might be a solution for this challenge. Also, the lack of competence and know-how of rebuilding and renovating machines were emphasized as an issue since in-house competence seldom no longer exist within manufacturing companies. In order to enable the circular production system, the competence of rebuilding and renovating the system

might not be required to exist within the focal manufacturing company, however, a close collaboration between manufacturing company and system supplier might be required.

Lastly, ***future legal demands*** were identified as a challenge towards circular production systems. Some participants were expecting stricter legal demands in manufacturing companies required to manage residual materials, but also an assigned amount of CO_2 emission for each manufacturing company. Hence, creating foundations for managing these added requirements on production systems to the need to establishing circularity in production were found key solution for managing future legal demands.

5 A Proposed Framework for Circular Production System

The proposed concept of circular production system is based on the 10R framework which was chosen due to its extensiveness compared to e.g., the 3R or 6R framework. Based on the inputs from the workshops, the Rs have been adapted into a production system domain (see Fig. 3). The Rs in the proposed framework is not sequenced in chronological order, nonetheless, similar to the product 10R framework, the main logic is based on the idea that an emphasis on the earlier Rs implies a higher degree of circularity.

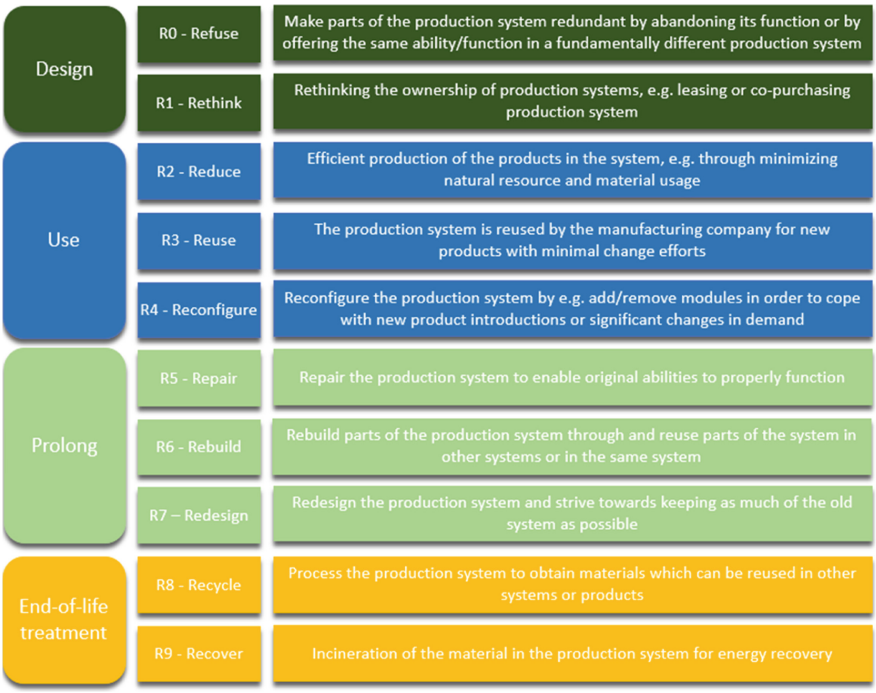


Fig. 3. Proposed conceptual framework for circular production systems.

Based on the Rs, four main phases have been developed: design, use, prolong and end-of-life treatment (EOL). These resembles the perceived main phases of developing circular production systems. The circular production system is based upon the notion that it is technologically possible to add and remove parts/modules within the system. A system which for instance cannot easily be reconfigured or repaired should not be recognized as a circular production system. Hence, when removing or adding parts/modules from the system, also these must be regarded in a circular manner. However, this is not regarded in the current conceptual framework for circular production system (Fig. 3) as it is solely from a systems perspective. In the design phase of a circular production system, Refusal of excessive production equipment and Rethinking the ownership are key activities. These set the foundation for how the circular production system is managed throughout its lifetime. In the use phase, the continuous reduction of natural resources and materials are sought for [8], but also in terms of e.g. energy, water and chemical usage [26]. In order to Reuse the production system, an inherent flexibility must exist, enabling a resource efficient change between products. When more extensive changes are required, e.g. when introducing a new product or fluctuating demands, Reconfiguration of the production system allows not needing to purchase a new system. Instead, solely adding or removing parts/modules in the system is sufficient to meet the changing requirements. Prolonging the production systems lifetime occurs when needed. In order to Repair and Rebuild the system, which are separated based on the extensiveness of the changes, the aforementioned changeability might be a key enabler. A system which cannot be easily repairable, might spur the action of purchasing a system new instead, in line with the linear economy. Lastly, in the cases where the production system cannot be used anymore, e.g. due to obsolescence or ineffectiveness, a well thought out end-of-life treatment should be executed. For instance, as with the 10R framework [8], the choice of recycling the materials in the production system is recognized a more circular option than energy recovery through incineration.

Although the 10R framework and the conceptual framework for circular production systems do not visualize the circularity itself, Fig. 4, resembles the circular approach covered in the framework based on the four phases. After iterative use and prolong phases, a new design phase will take place, as well as EOL-treatment of certain parts. Thereby, the resources which the production system itself consists of, and the resources used in the production system are maximally utilized, in accordance with the CE.

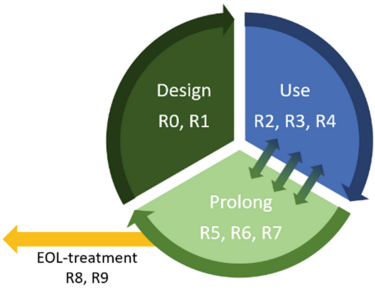


Fig. 4. The proposed circular production system illustrated as a near perpetual loop.

6 Conclusions and Future Research

Due to the ongoing climate crisis, manufacturing companies need to drastically reduce their negative environmental impact and adapt circularity. However, state-of-the-art CE frameworks are intended for traditional products whereas the end-user are consumers. As a counteract, this paper attempts at shedding light on the necessity to rethink how production systems are viewed. Through workshops with manufacturing company representatives, challenges towards circular production systems were identified and potential solutions for some were described. A conceptual framework covering the adaptation of state-of-the-art CE research to the domain of production systems has been introduced and described. The proposed circular production system is a system wherein both the resources used in the system, i.e. while using the system, but also the resources which the system consist of, is maximized in accordance with circularity. It is based on the key concept of maximining the utilization of raw materials and resources in order to aid in achieving a CE and thereby sustainable development.

A further continuation of research on the circular production system concept is recommended. The proposed conceptual framework solely includes environmental aspects, hence, further development of the framework could include investigating the sociological aspects. Future research might also include investigating enablers and barriers in circular production systems, empirically testing and validating the concept, as well as identifying best practices and guidelines for managing circular production systems.

References

1. Franco, M.A.: Circular economy at the micro level: a dynamic view of incumbents' struggles and challenges in the textile industry. J. Clean. Prod. **168**, 833–845 (2017)
2. Pitt, J., Heinemeyer, C.: Introducing ideas of a circular economy. In: Stables, K., Keirl, S., (eds.) Environment, Ethics and Cultures, Rotterdam: Sense Publishers, pp. 245–260 (2015)
3. Webster, K.: What might we say about a circular economy? some temptations to avoid if possible. World Futur. J. Gen. Evol. **69**(7–8), 542–554 (2013)
4. Bocken, N.M.P., de Pauw, I., Bakker, C., van der Grinten, B.: Product design and business model strategies for a circular economy. J. Ind. Prod. Eng. **33**(5), 308–320 (2016)
5. Bressanelli, G., Saccani, N., Perona, M., Baccanelli, I.: Towards circular economy in the household appliance industry: an overview of cases. Resources **9**(11), 1–23 (2020)
6. United States Environmental Protection Agency, "https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions (2021). Accessed 16 Feb 2021
7. Koren, Y., et al.: Reconfigurable manufacturing systems. CIRP Ann. - Manuf. Technol. **48**(2), 527–540 (1999)
8. Potting, J., Hekkert, M., Worrell, E., Hanemaaijer, A.: Circular economy: Measuring innovation in the product chain. Planbureau voor de Leefomgeving, 2544 (2017)
9. Jawahir, I.S., Badurdeen, F., Rouch, K.E.: Innovation in sustainable manufacturing education. In: 11th Global Conference Sustainable Manufacture, pp. 9–16 (2013)
10. Lieder, M., Rashid, A.: Towards circular economy implementation: a comprehensive review in context of manufacturing industry. J. Clean. Prod. **115**, 36–51 (2016)
11. Kirchherr, J., Reike, D., Hekkert, M.: Conceptualizing the circular economy: an analysis of 114 definitions. Resour. Conserv. Recycl. **127**, 221–232 (2017)
12. Geissdoerfer, M., Savaget, P., Bocken, N.M.P., Hultink, E.J.: The circular economy – a new sustainability paradigm? J. Clean. Prod. **143**, 757–768 (2017)

13. Wu, H.Q., Shi, Y., Xia, Q., Zhu, W.D.: Effectiveness of the policy of circular economy in China: A DEA-based analysis for the period of 11th five-year-plan. Resour. Conserv. Recycl. **83**, 163–175 (2014)
14. Badurdeen, F., Iyengar, D., Goldsby, T.J., Metta, H., Gupta, S., Jawahir, I.S.: Extending total life-cycle thinking to sustainable supply chain design. Int. J. Prod. Lifecycle Manag. **4**(1–3), 49–67 (2009)
15. Jawahir, I.S., Bradley, R.: Technological elements of circular economy and the principles of 6R-based closed-loop material flow in sustainable manufacturing. Procedia CIRP **40**, 103–108 (2016)
16. Cochran, D.S., Arinez, J.F., Duda, J.W., Linck, J.: A decomposition approach for manufacturing system design. J. Manuf. Syst. **20**(6), 371–389 (2001)
17. Bi, Z.M., Lang, S.Y.T., Shen, W., Wang, L.: Reconfigurable manufacturing systems: the state of the art. Int. J. Prod. Res. **46**(4), 967–992 (2008)
18. Wu, B.: Manufacturing Systems Design and Analysis: Context and Techniques. Chapman & Hall, London, U.K. (1994)
19. Elmaraghy, H.A., Wiendahl, H.P.: Changeability - an introduction. In: Elmaraghy, H.A., (ed.) Changeable and Reconfigurable Manufacturing Systems, pp. 3–24 London, U.K. (2009)
20. Koren, Y., Shpitalni, M.: Design of reconfigurable manufacturing systems. J. Manuf. Syst. **29**(4), 130–141 (2010)
21. Bi, Z.: Revisiting system paradigms from the viewpoint of manufacturing sustainability. Sustainability **3**(9), 1323–1340 (2011)
22. Garetti, M., Taisch, M.: Sustainable manufacturing: trends and research challenges. Prod. Plan. Control **23**(2–3), 83–104 (2012)
23. Asif, F.M.A.: Circular manufacturing systems: a development framework with analysis methods and tools for implementation. KTH Royal Institute of Technology [Doctoral dissertation] (2020)
24. Lieder, M.: From resource efficiency to resource conservation: Studies, developments and manufacturing systems. KTH Royal Institute of Technology [Doctoral dissertation] (2017)
25. Säfsten, K., Gustavsson, M.: Research methodology: for engineers and other problem-solvers, 1st ed. Lund: Studentlitteratur AB (2020)
26. Skärin, F., Rösiö, C., Andersen, A.-L.: An explorative study of circularity practices in swedish manufacturing companies. Sustainability **14**(12), 7246 (2022)

Facilitating Manufacturing System Development: Mapping Changeability Capabilities in Two Industrial Cases

Stefan Kjeldgaard(✉), Rasmus Andersen, Alessia Napoleone, Thomas Ditlev Brunoe, and Ann-Louise Andersen

Department of Materials and Production, Aalborg University, Aalborg, Denmark
stefank@mp.aau.dk

Abstract. In today's unpredictable markets, knowing where to focus development resources in manufacturing systems is critical for manufacturers. However, the effort required to completely assess a manufacturing system's ability to accommodate changes i.e., its changeability, is a resource intensive task whose outcome is not well understood in industry. This leads to companies being hesitant in applying extensive analysis methods. Therefore, this study presents the findings from applying an adapted and practitioner-oriented method for preliminary assessment of a manufacturing system's changeability, taking product and process characteristics as inputs. The method was applied in two manufacturing companies (Case A and Case B) which differ in scope and motivation. Focusing on assessing their current changeability, Case A found that the method provided a practical approach to map existing process capabilities while Case B found that no additional changeability was needed to accommodate planned product introductions. While the method proved to be versatile for different scopes and motivations, issues of capacity and scalability are not considered in the method. Furthermore, defining complete equipment capabilities is a challenge, especially for larger more complex systems, presenting the need for an efficient method to achieve this.

Keywords: Changeable manufacturing · Reconfigurability · Manufacturing systems development · Assessment · Case study

1 Introduction

Manufacturing companies are challenged with an increasingly turbulent and complex environment, due to reduced product life cycles, rising demand fluctuations [1], and increasing frequency of new product introductions [2]. These aspects, leading to challenging product evolution, have a significant impact on the evolution of manufacturing systems [3].

A solution to these challenges is developing changeable manufacturing systems; changeability allows accomplishing early and foresighted adjustments of the manufacturing processes, minimizing the effort required to change the structure of the manufacturing systems [4] and typically rely on the concepts of flexibility and reconfigurability.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 626–635, 2023.
https://doi.org/10.1007/978-3-031-16281-7_59

Furthermore, the increasing unpredictability around product evolution leads manufacturers to question, at different stages of a manufacturing system's life cycle, whether current capabilities are sufficiently changeable, or adaptations should be made. Furthermore, to ensure competitiveness and responsive reactions to sudden changes, manufacturers should be supported by practitioner-oriented methods to assess their changeability, so to rapidly identify where and how the system should be adapted [5–8].

The assessment of changeability is a complex task, where the requirements of current and new products should be considered in relation to the capabilities of the manufacturing system. With a focus on expected changes in product variety, Schou et al. [9] introduced a method to model and map relations between products and their related manufacturing system. The method is comprised of 10 sequential steps:

1. *"Establish a product (family) model representing the current variety.*
2. *Enrich the product model with attributes that are likely to change in the future.*
3. *Map the equipment and processes required to manufacture the product.*
4. *Determine the logical relations between components and processes.*
5. *For each component, evaluate whether changes to each component attribute would be possible to accommodate in the existing system by making hard or soft changes.*
6. *For each product attribute, determine what a change in the processes and equipment would imply for relevant processes and equipment.*
7. *Formalize relations between components and processes by formulating constraints/combination tables in declarative programming or in a configurator*
8. *Assign default values to variable in the current system setup*
9. *Running the configurator now allows inputting new values for variables and evaluating whether the new product family would be manufacturable in the system with or without changes, and which class of changes this would require in each process*
10. *If an attribute/capability is not present in the model, return to Step 2"* [9]

Nevertheless, the method proposed by Schou et al. [9] requires significant effort in collecting and analyzing relevant information, which as outlined earlier may prove an infeasible starting point for companies interested in assessing their changeability. Therefore, this study aims to address the following research question:

> *"What are the industrial insights from applying an adapted and practitioner-oriented version of the method proposed by Schou et al.?"*

2 Method

This section presents the practitioner-oriented method, based on Schou et al. [9], for assessment of manufacturing changeability. The method covers five steps as illustrated in Fig. 1.

Step 0 is delimitation to a specific area with respect to a family of products or parts and the related manufacturing processes that are assessed to be critical based on qualitative insights. This refers to an area which (i) currently constrains capabilities, (ii) has historically been prone to changes, or (iii) is expected to be changed in the future. The

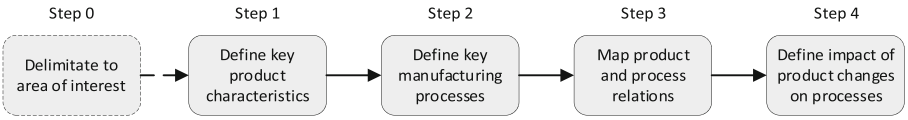


Fig. 1. The five steps and sequence of the proposed method for changeability assessment.

delimitation generates a practical starting point for small- and medium-sized enterprises (SMEs) with limited resources to identify immediate changeability issues.

Step 1 is to define the characteristics that determine if the product or part family can be manufactured in the delimited processes. They refer to capability requirements which (i) are critical for functionality changes of processes and their constituents, and (ii) are likely to change with new product or part introductions. For identified characteristics, the unit of measure is noted. For measures where the range of values is specified on a scale, such as dimensions, the minimum and maximum values are noted. However, for categorical values, such as material types, all are noted. For those measures which are common in the family, single values are stated.

Step 2 is to define the processes for manufacturing the delimited product or part family. It is done by mapping the processes and their interrelations. An aggregated mapping is proposed where processes are noted for their function e.g., grinding which can be decomposed to granular operations depending on the chosen delimitation.

Step 3 is to define the manufacturing processes which may be impacted by changes (i) between existing products/parts or (ii) from new product/part introductions. The critical characteristics and manufacturing processes are inputs for this step. A matrix is proposed to be applied for the mapping where qualitative data is entered. If a relation exists between a characteristic and a process, there is a potential impact of change on that process, and the relation is marked with "x". Although, if there is a higher degree of uncertainty, the relation is to be marked with "(x)" instead.

Step 4 is to define the impact of changing product/part characteristics in each manufacturing process. First, the processes' capability ranges are defined with respect to characteristics that carry an impact. This can be done by defining the processes' exact capability range, which is determined by their constraining constituents where data can be collected from their specifications. The range of existing characteristics can be used as a substitute, as they are within the processes' current capability range. Subsequently, the impact of changing between existing product characteristics to be processed within the current capability range is defined with respect to the required changeover time and cost. Moreover, the impact of converting to new product characteristics beyond the current capability range is defined in terms of lead-time and cost of new equipment.

2.1 Comparison of Methods

The focal method of this paper, relative to the method by Schou et al. [9], share a set of similar steps although they vary on aggregation level in terms of analyses and required data. The following paragraphs details common and varying elements between the methods' respective steps.

An initial divergence is present, as the focal method explicitly requires delimitation to a critical area. This contrasts to the broad coverage of the comparative method which increases the extent of analyses and required data. In step 1 the comparative method requires modelling of the (i) product variety of a family, (ii) the related break-down structure into components, and (iii) specified characteristic across. This contrasts to the focal method where critical characteristics are defined and specified for those that are subject to change, which does not include values expected by new products, which is required in the comparative method. Expectations or trends of future product/part characteristics are applied implicitly with regards to the fourth step of the focal method upon the comparison with the current capability range to derive required system changes. This is reasoned as the specifications are uncertain and most likely bound to change, making it suitable to remain unspecified until required by the analysis as (i) it is not needed until that point, and (ii) to minimize complexity and reiteration to reduce resources.

Subsequently, both methods propose mapping of related manufacturing processes, where the comparative method also decomposes the processes to equipment, which is delimited from in the focal method. Thereafter, Schou et al. [9] propose to map interrelations between the impacting characteristics and related processes, using similar matrices. However, the granular decomposition of the comparative method also requires mapping across lower levels e.g., components to equipment. This is not defined explicitly by the focal method, but rather proposed to be considered implicitly, to simplify the analyses.

However, in the pursuit of increased practical relevance, to generate the most relevant insights with minimum resource investment, a risk materializes regarding uncaptured critical knowledge. The latter relates to the fact that it is the constituents, such as equipment, within the processes that constrain their capabilities which requires definition of interrelations between components and constituents as proposed by the comparative method. However, such a comprehensive analysis can be daunting for SMEs entering the field with budget and time constraints. For this reason, the focal method sacrifices level of detail to increase the useability for practitioners, thereby facilitating rapid and easy identification of immediate critical areas where changeability issues arise. This is presumed to drive the motivation of SMEs to pursue further, more comprehensive, analysis.

The focal method defines the capability range in the fourth step, and the comparative method defines it in previous steps. One of the key aspects where they diverge is on how the change impact is measured: whether defined or specified. The comparative method defines it in terms of four changeability classes: (i) flexibility, (ii) soft changeability, (iii) changeoverability, and (iv) reconfigurability. In contrast, the focal method instead aims to specify the exact impact in terms of time or cost, whereas the latter is indirectly referenced by the required constituents e.g., new equipment.

These steps, regardless of method, are considered the core of changeability mapping Schou et al. [9]. The comparative method provides four additional steps which are excluded from the focal method due to the research objective. The objective does not require the formalization of relations and assignment of values in a configurator to automatically predict changes in manufacturing systems. Instead, the selected manual approach is deemed more practically applicable, and reasonable, for SMEs.

3 Findings

The method was applied in two industrial cases, hereafter respectively named Case A and Case B. Case A is a large sized company producing different kinds of metering solutions, while Case B is a small sized company producing sporting goods. In both cases, the manufactured product is not complex, this also affects the relatively low complexity in the resource domain. The companies have different reasons for interest in the method: being a large company, Case A has different manufacturing systems and resources currently dedicated to specific part families; however, there are specific part families, like the one analyzed in this paper, which share many commonalities with other part families, thus they expect that different manufacturing resources have similar capabilities. Knowing the capabilities of the existing resources and being able to compare them would bring value to the company due to demand fluctuations. Case B expects that a specific product family will undergo changes due to a rising demand for products that are bigger in size compared to the existing variety. Therefore, Case B used the method to investigate the impacts of changes in specific product parameters on the manufacturing resources.

Step 1. In both cases, key product characteristics were identified based on workshops with production engineers which resulted in the characteristics listed in Table 1. In both cases, nine different characteristics were identified and despite the products being very different in nature, several similarities of the identified characteristics are apparent. In both cases, the geometric dimensions of the product and material types used are important determinants of compatibility with production equipment. Other physical product properties are likewise identified such as part weight and product strength. Thus, while the study includes only two cases which are very different, several common characteristics are identified across the cases.

Table 1. Important product characteristics in relation to production equipment capabilities for the two cases.

Case A	Case B
• Size connection to peripheral	• Length
• Length	• Height
• Height	• Top surface material
• Depth	• Bottom surface material
• Internal diameter	• Fibre types
• Weight	• Fibre adhesive type
• Pallet interface	• Top surface adhesive type
• Communication interface	• Bottom surface adhesive type
• Material of top cover	• Strength

Step 2. Having identified the key product characteristics, Step 2 focuses on mapping the relevant production processes for the two cases. Figure 2 illustrates simplified process flows for the two cases. Despite the part in Case A being only a subassembly of the complete product, the production process comprises 12 relevant steps compared to the 7 steps of the product in Case B. Manual and automated processes are identified as relevant in both cases, while their share of manual processes differs substantially (17% for Case A and 57% for Case B).

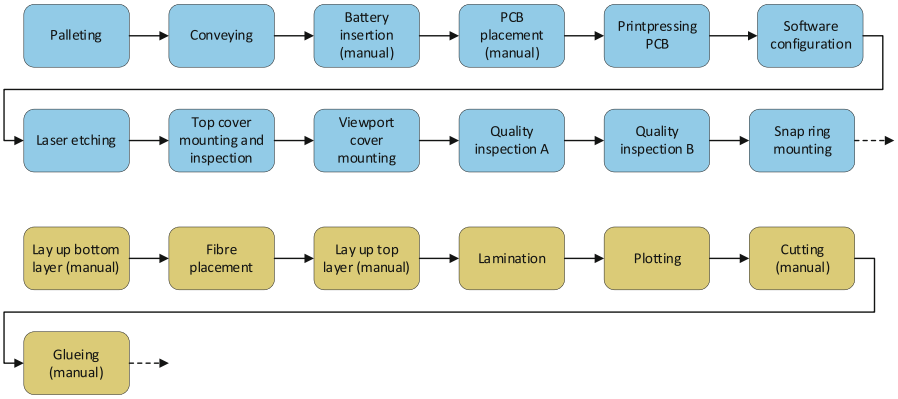


Fig. 2. Mapped production process steps for the two cases. Note: only processes within the scope of this study are mapped.

Step 3. After the product characteristics and production processes have been identified, relations between the two aspects are mapped in Step 3. Table 2 shows the resulting relations for Case A. It should be noted that while the method prescribes the use of product characteristics for this step, the case company opted for use of physical product components. The original approach, as described in Sect. 2 was applied in Case B (not shown).

The number of relations between product components/characteristics and related processes assist in the subsequent changeability mapping. Additionally, this information may also provide other helpful insights. For example, summing relations across columns indicates the criticality of changing a product characteristic in terms of the impact it may have across the production process and vice versa when summing across rows. Other aspects may be deduced from the information contained within these matrices, including how integrated the production process or equipment design is.

Step 4. The last element of the analysis, Step 4, focuses on the specific changeability of each production process from multiple aspects. Changeability may be analyzed according to current changeability only – as was done for Case A, see Table 3 – according to changeover and setup times, or according to future needs for changeability – the focus of the analysis in Case B, Table 4.

Table 2. Mapping of product-process relations for Case A. Note the usage of product components rather than product characteristics for this case.

	Production processes											
	1	2	3	4	5	6	7	8	9	10	11	12
Flow part	X	X	X		X	X		X	X	X	X	X
PCB	X			X	X	X		X			X	
Battery		X										
Top cover	X			X			X	X	X			
Glass	X			X					X			
O-ring	X			X					X			
Snap ring	X			X								X
Dry case		X			X	X			X			

Table 3. Changeability mapping example for Palleting process for Case A.

	Current changeability
Length	F: 105-1000 mm
Height	F: 0-210 mm
Diameter	D: 89,8 mm, R: 0-150 mm
Weight	F: 0-33 kg
O-ring diameter	D: 82 mm
Glass diameter	D: 79 mm

Based on the motivation of Case company A for performing this analysis, taking a descriptive focus on current changeability formed the scope of their project. The company found that the process of mapping product characteristics against production processes revealed that while some aspects of the production was very flexible, the constraining factor of their production lines was the diameter of the product, shown in Table 3 for the palleting process. This product characteristic relied on dedicated equipment throughout all production processes, and thus restricted the changeability of the production line substantially as changing diameter would require reconfiguration of most of the production processes. The insights gained from this analysis resulted in the discovery that production lines were typically designed around this product characteristic.

In Case company B, the motivation for performing the changeability mapping analysis was oriented more towards a predictive scope than a descriptive. The company was interested in investigating whether their production processes would support new product variants with changed product characteristics. Results from analyzing the production equipment revealed that the production processes employed by the company were very flexible with little need for extending flexibility to accommodate the planned product changes.

Table 4. Changeability mapping example for Bottom layer lay-up process for Case B.

	Current flexibility (F) and reconfigurability (R)	Impact on changeover (C) or setup (S) time	Required change for extending flexibility
Length	F: 0-19 m, R: 19-35 m	C: 2 x 3-4 hours (no downtime)	Reconfiguration
Height	F	-	-
Bottom surface material	F: 2 groups	C: No impact, S: 45 min	None (if within current groups)
Fiber adhesive type	1	-	-

4 Discussion

One of the trade-offs in the proposed method concerns the suitable aggregation level. An aggregated level of manufacturing processes is desired to simplify the analysis. However, it is the processes' constituents, such as the equipment, which constrain their capabilities [9, 10]. The proposed method does not explicitly support direct identification of what these equipment constraints are, only where they are located, although indirectly via the identification of the change impact. This indirect identification requires that the process is enabled by one-or-few types of equipment. However, in more complex processes, with an increasing amount, the link would be less apparent, where a mapping to capabilities of equipment would be suitable. This is also proposed in the method by Schou et al. [9] although both lacks industrial validation, where they also lack a specification of the exact capabilities and the impact with regards to changeability. Moreover, black-boxing of the equipment's influence carries additional implications as issues in material handling and storage equipment is not supported to be identified, due to the exclusion of inter-relationship between processes.

Another trade-off concerns the initial delimitation to a critical area, which does simplify the method, however, also requires a certain context. This critical area is presumed to be more apparent (i) when the break-down structure is less complex with respect to both domains, and (ii) within small- and medium-sized enterprises, where there is a more direct relationship between management and the workforce. However, for large-scale global manufacturing enterprises it would presumably be more beneficial to conduct a more exhaustive and comprehensive analysis, due to the complexity, and the uncertainty of where the critical areas arise, especially in fast moving and uncertain environments with frequent changes.

Further limitations, also apparent in Schou et al. [9], are that the requirements and capabilities regarding capacity and scalability hereof is excluded, as only functionality, changeoverability and convertibility hereof is in focus. This can be increasingly complex in environments where uncertainty in capacity and functionality requirements and capabilities are interlinked, where simulation could be beneficial to evaluate the best course of action with regards to impact mitigation.

Aforementioned limitations and trade-offs indicate that one-size does not fit all, with respect to the methods as objectives and contexts can vary to a high extent. Therefore,

further research is proposed on construction and industrial validation of methods which cater for limitations, to support industry in achieving the appropriate means for analyzing changeability, requirements, capabilities, mapping, impact, and mitigation hereof. If possible, a changeable method would be suitable.

5 Conclusion

This paper has presented a practitioner-oriented method for preliminary assessment of manufacturing system changeability based on domain expert input. Drawing inspiration from an existing method, this study reduced the scope and extensiveness of the original method by Schou et al. [9] to make it applicable for resource-efficient assessments of changeability in a practical setting. The proposed four-step method was applied in two manufacturing companies of different sizes with vastly different products. The output of the method provided sufficient basis for decision making in both cases, whether the focus was on assessing current changeability (Case A) or assessing needs for additional changeability for future product generations (Case B). It was furthermore found that the practicality of the applied method ensured continued engagement of both companies, which are now planning to apply the method to other products and manufacturing systems.

References

1. Echsler Minguillon, F., Schömer, J., Stricker, N., et al.: Planning for changeability and flexibility using a frequency perspective. CIRP Ann. **68**, 427–430 (2019). https://doi.org/10.1016/j.cirp.2019.03.006
2. Niroomand, I., Kuzgunkaya, O., Bulgak, A.A.: Impact of reconfiguration characteristics for capacity investment strategies in manufacturing systems. Int. J. Prod. Econ. **139**, 288–301 (2012). https://doi.org/10.1016/j.ijpe.2012.05.012
3. ElMaraghy, H., Monostori, L., Schuh, G., ElMaraghy, W.: Evolution and future of manufacturing systems. CIRP Ann. **70**, 635–658 (2021). https://doi.org/10.1016/j.cirp.2021.05.008
4. ElMaraghy, H., Moussa, M., ElMaraghy, W., Abbas, M.: Integrated product/system design and planning for new product family in a changeable learning factory. Procedia Manuf. **9**, 65–72 (2017). https://doi.org/10.1016/j.promfg.2017.04.008
5. Bortolini, M., Galizia, F.G., Mora, C.: Reconfigurable manufacturing systems: literature review and research trend. J. Manuf. Syst. **49**, 93–106 (2018). https://doi.org/10.1016/j.jmsy.2018.09.005
6. Zidi, S., Hamani, N., Kermad, L.: New metrics for measuring supply chain reconfigurability. J. Intell. Manuf. (2021).https://doi.org/10.1007/s10845-021-01798-9
7. Khanna, K., Kumar, R.: Reconfigurable manufacturing system: a state-of-the-art review. Benchmarking **26**, 2608–2635 (2019). https://doi.org/10.1108/BIJ-05-2018-0140
8. Singh, A., Gupta, S., Asjad, M., Gupta, P.: Reconfigurable manufacturing systems: journey and the road ahead. Int. J. Syst. Assur. Eng. Manage. **8**, 1849–1857 (2017). https://doi.org/10.1007/s13198-017-0610-z
9. Schou, C., Sørensen, D.G.H., Li, C., et al.: Determining manufacturing system changes based on new product specifications. J. Glob. Operar. Strateg. Sourcing (2020).https://doi.org/10.1108/JGOSS-10-2019-0060

10. Brunoe, T.D., Andersen, A.L., Sorensen, D.G.H., et al.: Integrated product-process modelling for platform-based co-development. Int. J. Prod. Res. **58**, 6185–6201 (2020). https://doi.org/10.1080/00207543.2019.1671628

Transfactory: Towards a New Technology-Human Manufacturing Co-evolution Framework

Fabrizio Bracco[1(✉)], Alessandro Arturo Bruzzone[1], and Emanuele Carpanzano[2]

[1] University of Genova, Genoa, Italy
fabrizio.bracco@unige.it
[2] University of Applied Sciences and Arts of Southern Switzerland, Manno, Switzerland

Abstract. The evolution of work organizational models and of technology poses an unprecedented set of challenges and opportunities for ensuring workers' well-being. After an overview of the evolution of these two aspects and the role of humans with their respect, we will frame the current and near-future developments according to the job demands-resources model, where humans' well-being depends on the good balance between job demands (workload, stress, risks, etc.) and job resources, both personal (resilience, optimism, etc.) and job-related (organizational and technological support). Adaptive automation and artificial intelligence have the potential to become job resources rather than job demands, but only if a proper design is set. Central to the design is the notion of plasticity, i.e., the capacity of an element to adapt to changes in its coupling with interacting elements. We will discuss two forms of plasticity in the relationship between humans and technology. Short-term plasticity is mainly based on the immediate adaptation of technology to human needs and performance. The long-term plasticity is based on the co-evolution of humans and technology, where changes concern not just the mere performance, but wider and long-lasting aspects like knowledge, culture, identities, approaches, job frameworks. Organization is crucial to foster this kind of co-evolution, towards a new framework, called transfactory, where human needs and values govern the overall systems evolution.

Keywords: Well-being · Plasticity · Adaptive automation · Artificial Intelligence · Organization

1 Introduction

The etymology of 'technology' comes from the Greek, 'τεχνολογία', by combining 'techne' (τέχνη) and 'logos' (λόγος), respectively art, skill, craft, and word, thought, principle, speech. Technology is currently pervading the humans' society. This keynote paper aims to assess possible strategies to integrate technology and humans within the manufacturing environment through a co-evolution approach able to meet psychological and manufacturing needs.

Several concepts that are nowadays used to characterize humans, industry, society, and complex systems were initially introduced in mechanical engineering, particularly to

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 636–645, 2023.
https://doi.org/10.1007/978-3-031-16281-7_60

study the mechanical properties of materials. Indeed stress, strain, elastic-plastic behavior, resilience are notions adopted by several disciplines ranging from psychology to economics. From the original mechanical context, governed by mechanistic, deterministic models, these notions are now used to characterize complex systems where humans and artifacts coexist, and indeterminism as well as chance are the rule.

Historically, manufacturing and human evolution are strongly related as machining capability has always influenced societies, their culture, determined military and economic power and in summary the dynamics of history.

Manufacturing depends on resources. The use of available natural resources, particularly force, provided in the beginning by animals, slaves, wind, rivers, evolved towards the concentration of power into artifacts, starting with the Watt steam engine, and successively the exploitation of functionalities, i.e. immaterial resources, granted initially by mechanical devices and nowadays by computers and Artificial Intelligence.

Currently, automation is a common factor in technology. According to Marx, "*automation is a process of absorption into the machine of the 'general productive forces of the social brain' such as 'knowledge and skills', which hence appear as an attribute of capital rather than as the product of social labor*".

In order to consider the actual industry into a comprehensive framework the relationship between human and artifact should be examined according to different features: technical, philosophical, organizational and psychological.

The relationship between humans, machines and nature within a holistic model is shown in Fig. 1 [1]. The homo habilis incorporates both theory and praxis to use nature and develop machines/systems in order to achieve a specific goal.

The development process of an artefact, simple machine or complex system, supposes the creator's will to satisfy needs that are congruent to the creator's values. This development process takes place within an historical context, characterized by scientific knowledge and social, aesthetic values. Considering the dynamics, the time required for the development process depends on the intensity of the involved resources; consequently, the resources concentration governs the acceleration of the development process and speeds up the industry and society evolution.

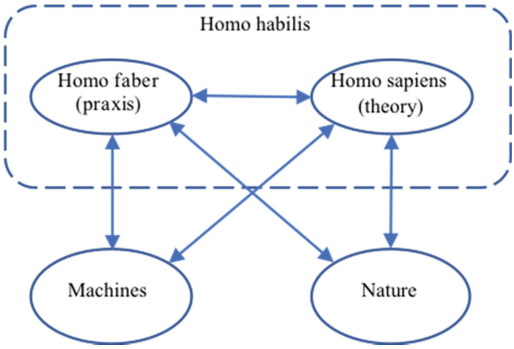


Fig. 1. Relationship between man, machines, nature [1].

Up to today, the interaction between human, machines and nature has been carried out according to the human will, that depends on its aesthetic values, and the available resources provided by nature or machines. In particular, considering the last three centuries western societies underwent essentially three main genuine industrial revolutions:

- The first spanned from 1770 to 1850: it is based on the concentration of power into steam engines, the development of railways with the consequent urbanization and reduction of distances;
- The second went from 1870 to 1910: it gave airplanes, car, electricity, telephone and radio communication;
- The third extends to 2000: it introduced and developed nuclear energy, nano and biotechnologies, computer and cognitive science.

The financial crisis of 2008 and its economic impact on industry, particularly on the automotive sector, lead, thanks to the German task group, to the INDUSTRIE 4.0 concept presented at the Hannover Industrial Fair in 2011 [2, 3]. The new revolution focuses heavily on interconnectivity, automation, machine learning, and real-time data; it requires Key Enabling Technologies (KETs) specifically, nanotechnologies, photonics, advanced materials, biotechnology, and advanced manufacturing. In 2015 the term Industry 4.0 was widely promoted by Klaus Schwab, the World Economic Forum Founder and Executive Chairman, emphasising that the changes provided by Industry 4.0 are not limited to efficiency improvements, but would have a significant impact on industrial capitalism [4].

With the 5th Science and Technology Basic Plan, the Japan Prime Minister Shinzo Abe proposed Society 5.0, a human-centered society that balances economic advancement with the resolution of social problems by a system that highly integrates cyberspace and physical space [5]. Currently the available technologies are not limited to the physical domain, but by using Artificial Intelligence extend their capability to the cognitive process previously specific of humans.

In particular this paper addresses the aspects related to the introduction of Artificial Intelligence in the relationship between human, machine, and nature. Specifically, starting from the concept of elasticity and plasticity, the paper discusses the properties of a recently introduced safety model, the well-being bowl, and its use to implement an advanced organization for future manufacturing systems.

2 The Human-Technology Reference Manufacturing Framework

2.1 The Human in the Loop Factory Reference Model

Digitalization is strongly changing industrial systems and related value chains. In order to deal with continuous product innovation dynamics, production processes have to increase flexibility and embedded intelligence at different levels, and to integrate advanced control and automation systems [6]. Moreover, in the manufacturing context advanced technologies such as collaborative robots, exoskeletons, and wearable sensors are spreading with the aim of assisting workers and improving their working conditions.

As a matter of fact, it is becoming of major relevance to balance such technological innovations with organizational aspects and human factors, to reduce human errors, increase productivity, and enhance safety and comfort of industry workers [7, 8]. Specifically, psychological and physiological principles have to be considered in the design of industrial plants, with a specific focus on the interaction between the human and the digital control and automation system [9, 10].

A conceptual and modelling framework that supports a seamless adoption of a human-in-the-loop adaptive automation approach has been introduced in [10]. By optimizing the capabilities residing in the human dimension of the factory and the digital automation's full potential within a unique coherent framework, it becomes possible to fully enhance industrial productivity and human workers safety and well-being at the same time [11, 12].

Figure 2 shows the proposed human-in-the-loop factory control and automation system reference model. Such a framework includes the human as a fully integrated part of the whole process, considering his two dimensions as active operator and decision maker. The human acts at operational level taking directly part to the production process, as well as at decision making level, in synergy with supervision and execution planning tools of the industrial automation systems.

Within the proposed digital-cognitive automation framework, goals specifically aimed at enhancing working conditions operate in parallel with traditional performance targets. Consequently, a stream of physiological measures is constantly monitored to detect in real time any deviations from personalized safe patterns, and propose actions aimed at mitigating the cognitive and physical demand that the worker is experiencing [13]. Moreover, reconfigurable automation policies that apply in the distributed automation structure [14] explore and possibly eliminate the sources of cognitive and physical gaps such as skill mismatching and alienating duties.

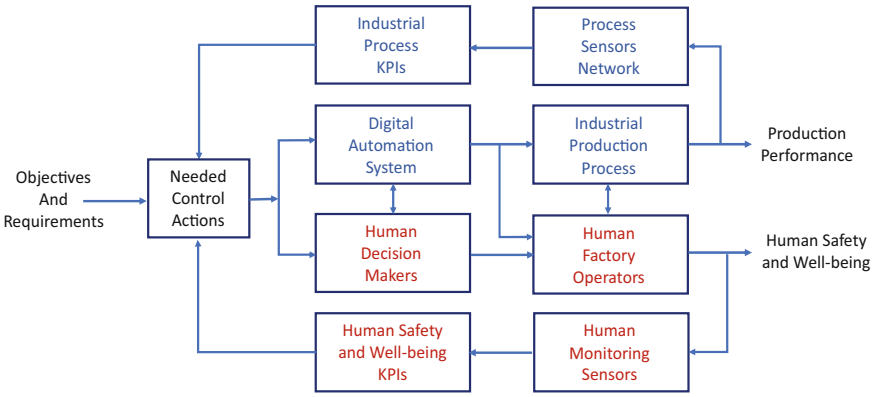


Fig. 2. Framework for human-in-the-loop factory adaptive control.

2.2 The Human Well-Being Bowl

In order to represent operators' safety and well-being, a systemic model has to be introduced that includes the cognitive and physical aspects of operators interacting with the industrial control and automation system [15]. Relevant facets to be considered are: the human operator; the team; the organization; the physical environment; the social environment; the tools (including technology, control and automation systems); the rules and procedures; as well as the execution tasks [10]. These features can be regarded as interacting, like fragments of a bowl that dynamically move to fill in the gaps that may arise at their borders (Fig. 3).

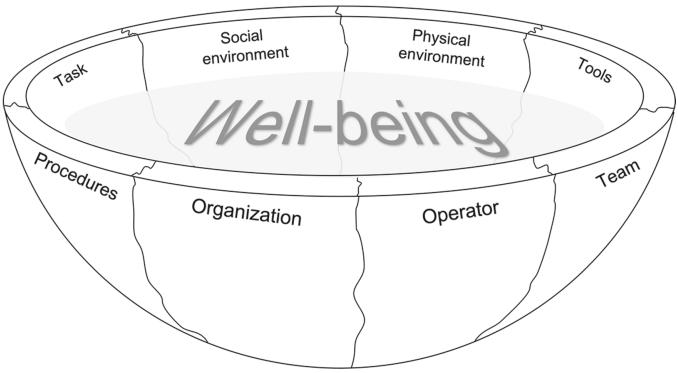


Fig. 3. The human well-being bowl.

The water in the bowl represents the operators' well-being, whose optimal level depends on the dynamic interaction among the main aspects of a system; well-being level is determined by the gaps that the fragments create at their borders. Sometimes an action aimed at increasing well-being could consider only one fragment; but changing just one part could lead to breaking the bowl if the other elements do not adapt to it.

The design of digital industrial automation systems should be guided by considering this model to promote harmonization of the fragments' behaviour. In particular, adaptive automation could provide a flexible fragment that copes with the inherent variability of the production system coming from human and organizational factors, process changes and productivity needs [6, 11]. The constant adaptation has to enable the system to fill the gaps ensuring operators' overall safety and well-being.

3 Human and Machine Co-evolution: Plasticity and Elasticity

During the manufacturing system life cycle the related technological, organizational, and human aspects will continuously evolve; therefore, a unique coherent and consistent reference framework has to be introduced to properly manage such a technology-human co-evolution.

In fact, as production needs will change according to market dynamics and product innovation, the manufacturing process will integrate new technologies and its control and automation system will be reconfigured whenever necessary.

As a consequence, the human workers will be faced with continuously changing jobs demands, and a proper balance between job and personal resources has to be guaranteed to achieve production objectives while guaranteeing workers well-being at the same time (Fig. 4, [16]). In particular, whenever the manufacturing systems is reconfigured, a new balance between the capabilities provided by the machines and the automation systems and the activities of the human workers has to be defined within a harmonised co-evolution framework, guaranteeing production effectiveness and human well-being at the same time.

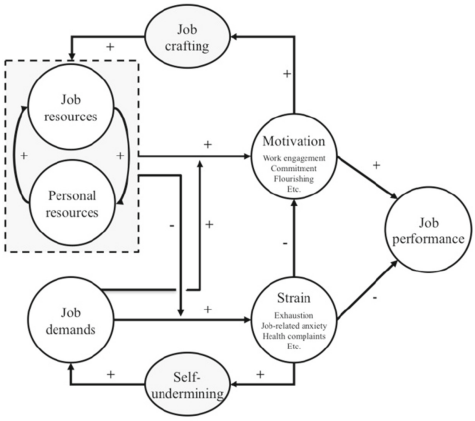


Fig. 4. The job-demands resources model [16].

According to the job demands-resources model, workers' performance and well-being is the result of a dynamic balance between resources and demands. The resources could be related to the job. They are organizational, social, and contextual aspects that are functional to the achievement of work objectives, reduce work demands and their psychological consequences, and stimulate personal growth. The resources are also present in the person, they are personal characteristics that allow the workers to have a positive evaluation of their own skills and to effectively manage the interaction with the environment (e.g., self-efficacy, optimism, resilience, etc.). The virtuous integration of resources leads to motivation and to job crafting, i.e., material change (intervention on the type of activity, the amount of work, the content of tasks), and cognitive changes (interpretation and meaning given to activity) that help the workers to dynamically adapt to the job demands. Job crafting can also act on the demands of the task, trying to increase the positive challenges and reducing the negative load.

On the other hand, there are job demands, i.e., organizational, social, technological and physical aspects that require an effort and therefore a cost for workers, on the physical and/or psychological level. Job demands may have a positive impact on performance if they are perceived as positively challenging, otherwise they are called hindrance

demands, and lead to self-undermining behaviors (e.g., dysfunctional reactions to stress like irritability, mistakes, conflicts, etc.), and eventually to strain and a decrease in safety, well-being, and performance.

The reconfiguration of the manufacturing system by means of changes in automation and, in general, in technology, could become a job resource or a job demand, depending on its integration with the human element [17].

Technology could become a demand if its effects on workers are not considered. For instance, when automation performs the most repetitive tasks, workers are left to cope with more cognitively demanding tasks like dealing with new or complex information and engaging in vigilance and maintenance tasks. Again, the introduction of a new technology requires new knowledge, the workers cannot simply forget about their previous habits, skills, and practices. Sometimes it is not just a matter of training, but also of cultural change in the perception of one's own professional identity.

Technology could become a resource if it designed to become an effective support of workers and if allows job crafting behaviors, where the agency of the human element (i.e., the capacity to act) is valued and facilitated by the organizational context. The impact (positive or negative) of technology over human well-being could be understood, therefore, not only by considering how technology is well matched with the workers. This perspective is focusing just on the interface between two of the several pieces of the well-being bowl: i.e., the human and the tools. We propose to investigate how all the pieces of the system interact after the introduction of technological changes, because they may affect workers' well-being not directly, but indirectly, through the impact on procedures, or organizational culture, or team dynamics.

Job resources and personal resources, therefore, should be able to co-evolve, to interact and change synergically. Technology, as a job resource, and workers, must be analyzed in their dynamic interaction, and not just as static matching. The well-being bowl model poses that the system is constantly stressed by changes in one or more of its elements (from the stakeholders needs and requests, to technological evolution, from social dynamics to regulatory developments). To cope with this kind of changes, the system should have some degrees of flexibility, i.e., the capacity to modify in accordance with perturbations. The several elements of the bowl, however, have several degrees of flexibility and it important to take them into account. For example, expecting workers to be able to suddenly change work habits is unrealistic, while reprogramming is relatively quick for machines. On the other hand, humans can adopt an exceptional violation to cope with an unpredicted situation, choosing the best option given the circumstances, while machines could be less sensitive to the specific situating and be more constrained by procedural routines.

Borrowing terms from material science, we could say that the human and machine co-evolution should be based both on elasticity and plasticity. Elasticity is the capacity of the system to change its setting to adapt to some demand or stress, but to recover its previous state when the demand or stress is over. Plasticity, on the other hand, is the capacity of the system to change and evolve maintaining its setting according to the new conditions; plasticity modifies the cognitive model by capitalizing the experience gained during the effort to face new demand or stress. Humans are particularly elastic; they are constantly able to adapt rules and procedures to the here and now. Technology

is starting to show some degrees of elasticity, but not as developed as the humans. On the other hand, technology could be more plastic; it can change and evolve to cope with new demands and can maintain its new setting over time. On the contrary, humans may be more reluctant to definitive changes, especially if they are perceived as threatening their professional identity, their habits, and practices. Or even if they require some intensive training, some cultural or organizational change.

Adaptive automation and artificial intelligence provide an opportunity to reinforce the co-evolution between humans and machines because they enable not only the short-term plasticity (i.e., the immediate adaptation of technology to human needs and performance), but also the long-term plasticity (the co-evolution of humans and technology, where changes concern not just the mere performance, but wider and long-lasting aspects like knowledge, culture, identities, approaches, job frameworks).

The organization has a crucial role of enabling this co-evolution based on plasticity, especially allowing workers to engage in job crafting, giving them the tools to develop their professionality, set goals, define strategies, open to new challenges. The adaptability of automation and the "intelligence" of technology should harmonically interact with these changes, helping the system not only to adapt to the environment, but to pursue human safety and well-being.

In biology, evolution is based on variability. Many of these changes led to dead ends, while some of them occurred to be successful and became a further step of life on Earth. In the organizational context, variability is an intrinsic property of many elements of the system, especially those related to humans (workers, teams, decision makers, stakeholders, etc.). In Resilience Engineering literature [18], human performance is described as intrinsically variable and should not be compared to a machine-like performance. There will always be a gap between the "work-as-imagined" and the "work-as-done" [19]. The solution is not to forcefully flatten the human variability ("work-as-done") to procedures and rules ("work-as-imagined"). The resilient organization is able to understand and monitor this gap, keeping it within a safe range, because the gap may be source of new developments.

This variability, like in biological evolution, could lead to a better adaptation, an innovative solution, or to useless or even harmful outcomes. According to this perspective, human variability should be treasured, as an evolution driver. However, understanding which kind of changes and how much variability could be beneficial for the whole system is not clear and is a critical challenge. In this paper, we claim that human and technology co-evolution could be based on this capacity to monitor the gap between "work-as-imagined" and the "work-as-done" and, thanks to artificial intelligence and adaptive automation, to keep it within a safe range.

4 Open Issues: Towards Transfactory

Organization is crucial to foster this kind of co-evolution. We need a new framework where human needs and values govern the overall system evolution thanks to a virtuous integration with technology [20].

Ryan and Deci [21] proposed one of the most accepted theories of human needs and motivation, and it has been applied to many contexts of human activities, also that of

organizational well-being [22]. Human motivation and well-being are favored when the work conditions meet the three basic needs of autonomy, belongingness, and competence. They all could be satisfied by means of a virtuous interaction with technology. The need for autonomy drives towards situations where the workers can experience some freedom in their choices and some agency (think about job crafting, for instance). Adaptive automation does not force the workers to some specific pace or task, it lets them the autonomy to set the conditions of work. The need for belongingness is related to the propensity to feel connected to others, to feel like a member of a team, a community of workers. Technology should not isolate workers, it should let them perceive their work as the result of a collective endeavor. The need for competence is the desire to feel that the work challenges are affordable, that the worker can master professional practices and bring about desired outcomes. Technology can help workers to prevent the feeling of being overwhelmed by the task complexity or task load, becoming a proper job resource.

The co-evolution of technology and humans can lead to the satisfaction of human needs, for the sake of motivation, well-being, and happiness. Within this perspective, technology becomes a tool for human self-actualization and fulfilment of needs and values. A similar approach has been proposed by transhumanism, a cultural movement that affirms technological and scientific modernity to obtain, in favor of the human species, physical and physiological benefits. It aims at the improvement of health and psychological well-being, with enhancement of intellectual and social abilities thanks to science and technology [23]. We therefore propose the term "transfactory" as the next step in the development of sociotechnical production systems. The co-evolution of humans and technology can counterbalance job demands with the virtuous coupling of personal and technological resources. In this perspective, the transfactory becomes a resilient ecosystem which is able not only to adapt to perturbations, but to take advantage of variability and to turn it into innovation.

References

1. Bruzzone, A.A., D'Addona, D.M.: Mind, machines and manufacturing: a philosophical essay on machining. Procedia CIRP **79**, 580–585 (2019)
2. Kagermann, H., Lukas, W.D., Wahlster, W.: Industrie 4.0: Mit dem Internet der Dinge auf dem Weg zur 4. industriellen Revolution, VDI news, 2011, 1 April 2011 (2011)
3. Kagermann, H., Wahlster, W.D., Helbig, J.: Recommendations for Implementing the Strategic Initiative Industrie 4.0, Frankfurt (2013). https://doi.org/10.1016/0375-9601(80)90605-2
4. Schwab, K.: The Fourth Industrial Revolution. Currency (2017)
5. Önday, Ö.: Society 50 - its historical logic and its structural development. J. Sci. Rep. **2**(1), 32–42 (2020). https://doi.org/10.5281/zenodo.3731954
6. Carpanzano, E., Jovane, F.: Advanced automation solutions for future adaptive factories. CIRP Ann. Manuf. Technol. **56**(1), 435–438 (2007)
7. Endsley, M.R., Kiris, E.O.: The out-of-the-loop performance problem and level of control in automation. Hum. Factors **37**(2), 381–394 (1995)
8. Lee, J.D., Seppelt, B.D.: Human factors and ergonomics in automation design, 3rd edn. In: Handbook of Human Factors and Ergonomics, pp.1570–1596 (2006)
9. Onnasch, L., Wickens, C.D., Li, H., Manzey, D.: Human performance consequences of stages and levels of automation: an integrated meta-analysis. Hum. Factors **56**(3), 476–488 (2014)

10. D'Addona, D.M., Bracco, F., Bettoni, A., Nishino, N., Carpanzano, E., Bruzzone, A.: Adaptive automation and human factors in manufacturing: an experimental assessment for a cognitive approach. CIRP Ann. **67**(1), 455–458 (2018)
11. Valente, A., Mazzolini, M., Carpanzano, E.: An approach to design and develop reconfigurable control software for highly automated production systems. Int. J. Comput. Integr. Manuf. **28**(3), 321–336 (2015)
12. May, G., Taisch, M., Bettoni, A., Maghazei, O., Matarazzo, A., Stahl, B.: A new human-centric factory model. Procedia CIRP **26**, 103–108 (2015)
13. Bettoni, A., et al.: Mutualistic and adaptive human-machine collaboration based on machine learning in an injection moulding manufacturing line. In: Procedia CIRP – 2020 53rd CIRP Conference on Manufacturing Systems, vol. 93, pp. 395–400. Elsevier (2020)
14. Carpanzano, E., Bettoni, A., Julier, S., Costa, J.C., Oliveira, M.: Connecting humans to the loop of digitized factories' automation systems. In: Ni, J., Majstorovic, V.D., Djurdjanovic, D. (eds.) AMP 2018. LNME, pp. 180–193. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-89563-5_14
15. Bracco, F., Brunel, C., Piccinno, T.F., Sedaoui, A., Tazi, D.: Promoting safety through resilient organization managers: different ways of being resilient. In: Proceedings of the 3rd SAFRA Symposium, Paris, 9–10 February 2015 (2015)
16. Bakker, A.B., Demerouti, E.: Multiple levels in job demands-resources theory: Implications for employee well-being and performance. In: Diener, E., Oishi, S., Tay, L. (eds.) Handbook of Wellbeing. DEF Publishers, Salt Lake City, UT (2018)
17. Demerouti, E.: Turn digitalization and automation to a job resource. Appl. Psychol. (2020)
18. Hollnagel, E.: The four cornerstones of resilience engineering. In: Resilience Engineering Perspectives, vol. 2, pp. 139–156. CRC Press (2016)
19. Hollnagel, E.: Why is work-as-imagined different from work-as-done? In: Resilient Health Care, vol. 2, pp. 279–294. CRC Press (2017)
20. Ueda, K., Takenaka, T., Váncza, J., Monostori, L.: Value creation and decision-making in sustainable society. CIRP Ann. **58**(2), 681–700 (2009). https://doi.org/10.1016/j.cirp.2009.09.010
21. Ryan, R.M., Deci, E.L.: The darker and brighter sides of human existence: basic psychological needs as a unifying concept. Psychol. Inq. **11**, 319–338 (2000)
22. Van den Broeck, A., et al.: Self-determination theory: a theoretical and empirical overview in occupational health psychology. In: Occupational Health Psychology: European Perspectives on Research, Education, and Practice, vol. 3, pp. 63–88. Nottingham University Press, Nottingham (2008)
23. Manzocco, R.: Transhumanism. Engineering the Human Condition. Springer, Suiza (2019).https://doi.org/10.1007/978-3-030-04958-4

Sustainability Comparison of a New Reconfigurable Machine Tool and a Conventional Milling Machine

Alessandro Arturo Bruzzone[1(✉)], Alessandra Ferrari[1], and Alessia Napoleone[2]

[1] DIME, University of Genova, Genova, Italy
Alessandro.Bruzzone@unige.it
[2] Department of Materials and Production, Aalborg University, Aalborg, Denmark

Abstract. Since few decades sustainability is an increasingly emerging and relevant topic supported nowadays by international policies. Many restrictions and rules have been introduced in order to guarantee a sustainable development and meet most of human society's future goals. Although sustainability is a very wide topic that concerns different subjects, this paper, focusing on manufacturing companies and on the importance of reconfigurable machine tools, uses the term "sustainability" considering both an environmental and industrial point of view. In particular the evaluation of the sustainability of a new Reconfigurable Machine Tool (RMT) is proposed and compared to a conventional milling machine tool. The Life Cycle Assessment (LCA) method has been used for the sustainability evaluation; it has been implemented through the open-source software OpenLCA. The analysis shows that the new Reconfigurable Machine Tool has lower pollutant emissions and environmental impact values are lower than the conventional milling machine tool.

Keywords: Sustainability · Reconfigurable machine tool · Milling

1 Introduction

Reliable scientific information is fundamental to drive sound decisions especially when the decisions impact on mankind welfare. In the last decades environmental sustainability and energy availability emerged as critical topics [1]; in particular global warming and sustainable development are increasingly key themes. In 1997 the Kyoto Protocol following the United Nations Framework Convention on Climate Change, required to the industrialized countries a transition to limit and reduce greenhouse gases emissions in accordance with agreed individual targets [2]. Even though recent studies provide contrary scientific evidence on climate change [3], a conscious industrial policy must address the effect of manufacturing on the environment.

Nowadays sustainability should drive the transition to a green economy, defined as "an economy that results in improved human well-being and reduced inequalities over the long term, while not exposing future generations to significant environmental risks

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 646–656, 2023.
https://doi.org/10.1007/978-3-031-16281-7_61

and ecological scarcities" [4]. Sustainability concerns different fields that range from environment to economics and social science [5, 6].

Since this paper focuses on manufacturing companies, the term sustainability is related to the intersection between the environmental and industrial domains. Industrial sustainability entails the optimization of the entire product life cycle, ranging from the raw material to the component, the product, the product fallen in disuse and its final disposal, considering factors such as resources, energy and assets.

Until recently, manufacturing technology evolved to achieve high performance at low cost, leaving environmental issues secondarily. However, environmental factors are becoming an emerging dimension in manufacturing processes due to increasingly stringent regulations and standards that impact on the production wastes, on the product life cycle, and furthermore drive the consumers' opinion toward green products.

Manufacturing companies are also facing frequent and unpredictable changes in the market that require a competitive and adaptable approach by using production paradigms that respond cost-effectively and quickly to uncertainty. To this end, Reconfigurable Machine Tools (RMT) are able to change their configuration providing the type of machining needed by specific manufacturing operations with a few and quick steps, these could provide the adaptability required by market demand variation.

The study evaluates the sustainability of a new RMT, against a conventional milling machine tool through a Life Cycle Assessment (LCA).

2 Framework of the Study

The introduction of a new technology requires a sustainability study in order to evaluate its impact in comparison to the current used technology. Nowadays there are no studies which compare the sustainability of a new RMT with a conventional milling machine. The analysis requires tools provided by Industrial sustainability, Life Cycle Assessment as well as the characteristics of the new RMT.

2.1 Industrial Sustainability

The term sustainability refers generally to the capacity for the biosphere and human civilization to coexist. At the base of sustainability lies the concept of sustainable development. In 1987, the United Nations World Commission on Environment and Development introduced the definition of "sustainable development" as "the development that meets the needs of the present without compromising the ability of future generations to meet their own needs". This definition has been widely adopted, despite it does not provide any specification of what kind of needs are concerned.

Nowadays the theme concerning the sustainability and the sustainable development, is becoming more and more relevant and it represents the essence of most of the future aims of our society. According to this, in fact, in 2015 United Nations Member States subscribed the 2030 Agenda, an action program for people, the planet and prosperity, which incorporates 17 Sustainable Development Goals into a major action program for a total of 169 'targets' or goals to be achieved by 2030 [7].

Two more accurate and specific definitions, one focused on the company environment and the other one on industrial sustainability must be introduced. As far as the first definition is concerned, John Elkington [8] added the concept of sustainability in a company context by suggesting three dimensions of sustainability, specifically social, environmental and economic, reflected in three basic lines as people, planet and profit, which are features that a company must guarantee to be sustainable (Fig. 1).

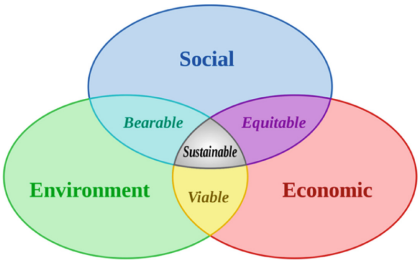


Fig. 1. Diagram of sustainable development, resulting from the intersection of the three constituent parts [8].

The industrial sustainability involves the optimization of the entire material cycle, starting from the raw material to obtain the component, the product, the product fallen in disuse and its final disposal, considering factors such as resources, energy and assets. Furthermore, industrial sustainability includes research on alternative energies, the use of new materials, new technologies and new technological system, but also legislative, social, economic and management aspects. The industrial sustainability includes several principles, such as:

- Commonly used products, as well as processes and services, may produce residues, but not waste.
- Each process, each product and each technological system should be adequately planned so that they are easily suitable for any innovations in order to protect the environment.
- Industries should receive most of the materials, subsequently used, by recycling systems, rather than always using raw materials.
- The materials used should be less toxic as possible for the purpose for which they are employed.
- Industries should make minimal use of materials and energy in product, process or service manufacturing.
- Each product should be designed in the best way in order to use it for creating useful products after its current use.
- Strong interactions should be created among suppliers, customers, and representatives of other industries in order to develop good ways of cooperation, to minimize packaging as much as possible and to recycle and reuse materials.

These principles are the basis of any process or product design that considers ecological and sustainable aspects, and represent a guideline to choose materials, energy

sources and industrial strategies, concerning infrastructures and production flows. From this perspective, some efforts have been made to implement industrial sustainable applications, such as design for environment methods in the electronics industry, integrated pest management or life-cycle assessment methodologies in mechanical and common object industries. In addition to the principles listed above, criteria have been defined, aiming at encouraging sustainable designs of manufacturing processes. According to this, in fact, the manufacturing system life cycle could be outlined in five phases which are the choice of raw materials, the production, the distribution and packaging, the use and the end-of-life cycle.

2.2 LCA Methodology

Nowadays the concept of Sustainable Development is supported by the innovative Life Cycle Thinking philosophy, which, unlike in the past, considers a product as a set of operations, input and output flows of materials and the forms of energy associated with all the steps of its life cycle, from design to dismantling and recovery or final disposal.

The Life Cycle Assessment (LCA) methodology allows a complete study of the product considering all the involved processes related to its entire life cycle [9]. The concept of life-cycle assessment first emerged in the late 1960's but did not receive much attention until the mid-1980's. In 1994 the International Standards Organization (ISO) began developing standards for the LCA as part of its 14000 series standards on environmental management. The standards address both the technical details and conceptual organization of LCA; they are:

- ISO 14040-A standard on principles and framework,
- ISO 14041-A standard on goal and scope definition and inventory analysis,
- ISO 14042-A standard on life-cycle impact assessment,
- ISO 14043-A standard on life-cycle interpretation.

Several LCA methodologies follow the LCA framework defined in ISO 14040, involving an inventory similar to that described in ISO 14041, and assessment of impacts to some degree as described in ISO 14042, while a smaller number takes on the normalization and weighting also discussed in ISO 14042. However, methodologies based on the ISO standards may differ greatly, since the ISO standards allow flexibility to customize characterization and normalization factors and weighting methods to suit the values and conditions of a particular location or sector.

The LCA methodology is implemented through several steps: the goal and scope definition, inventory analysis, impact assessment, and the interpretation of results. LCA has some limitations since it depends on standard databases that refer to general figures independently from the examined specific process. LCA can be integrated with an exergetic analysis considering a thermodynamic approach. In fact, the energetic analysis, which is the one related to LCA, measures the energy required directly and indirectly to produce a specific commodity or service without considering the loss of energy quality. The exergetic analysis examines instead the change in efficiency in a more practical way [10]. Other assessments can rely on a thermoeconomic analysis based on the flow-fund model of production [11].

LCA has been carried out by using OpenLCA [12], a free open-source, professional LCA and footprint software with a broad range of features and many available databases, created by GreenDelta since 2006. OpenLCA has a powerful plug-in structure which allows for easy extensions and modifications. Actually, to assess sustainability myriad tools and methods are available; sometimes these tools are competing and often contradicting; the OpenLCA software framework is designed to cope with this complexity by a proper selection of reference databases.

In this study two databases have been employed simultaneously: the ELCD 3.2 and the Environmental Footprints database (EF). The ELCD 3.2 contains nine new processes and, in total, it shows 503 processes for sectors such as energy, transport and materials production [13]. The EF database is designed to support the use of Product Environmental Footprints (PEF) category rules and organization environmental footprint sector rules; it contains also secondary life cycle inventory datasets intended to be compliant with the EF method, and a related EF impact assessment method. Objective of PEF is to develop a common methodology on the quantitative assessment of environmental impacts of products, in order to support their assessment and labelling.

LCA methodology requires to identify the aspects related to the activities, products and services interacting with the environment. The three main considered categories concern the materials of the product, the use of the product, the transport. Information about the materials, consumable, energy consumption, waste, disposal of consumables is required. Once the environmental aspects are known, these must be transformed into impacts which can be assessed with or without Eco-Indicators providing respectively a quantitative or a qualitative analysis. The Dutch Ecoindicator '99, presents five main categories: materials, processes, transport, use, disposal [14]. The present analysis considers water consumption, energy consumption, atmospheric emissions, fossils and the climate change connected to the fossil used resources.

2.3 The New Reconfigurable Machine Tools

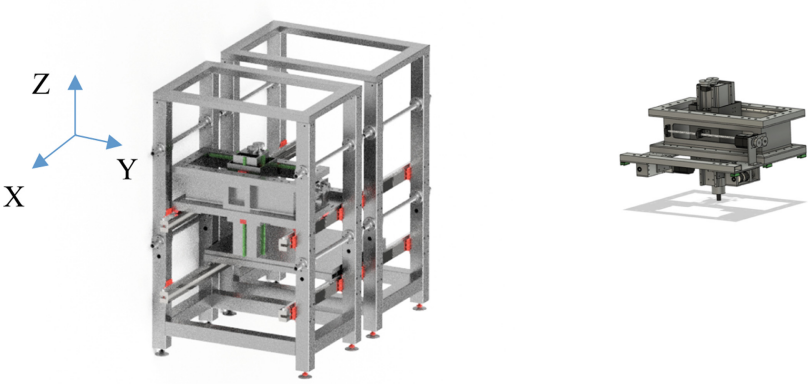
A Reconfigurable Machine Tool (RMT) is a machine tool that is able to change its configuration based on the type of machining required in a certain period, which might vary due to changed market requirements and/or technological advances, with few, simple and quick steps [15]. The possibility to quickly reconfigure these machine tools depends on their modularity. In general, modularity is the degree to which system's components may be separated and recombined, often with the benefit of flexibility and variety in use. Modularity is used in order to reduce complexity by breaking a system into interdepend and independent subsystems, assembly or components; this concept is at the base of the reconfigurable machine tools. According to this approach, it is possible to increase machine tools efficiency and application scope through the development of machine tools modules, which can be quickly assembled and disassembled. Furthermore, from an industrial sustainability perspective these machines are interesting because adapting the same machine tool to different machining processes can reduce the total energy consumption during the manufacturing process, the need for more space in the factory and the cost per part.

In this study RMTs are supposed to perform processes such as turning, milling, drilling, grinding or combinations of these operations and achieve their main goal i.e., to

manage various changes in the product or parts that must be manufactured. The possible product variations concern: part size, 3D geometry and complexity of a part, volume and speed of production, required processes, accuracy requirements in terms of geometric accuracy, surface quality and material properties.

a) b) c)

Fig. 2. New RMT: a) bed module, b) three beds line, c) base cross-table.



a) b)

Fig. 3. RMT two beds line (a) hosting milling spindle cross-table (b) with X, Y and Z axes.

The new RMT considered in this study can change the configuration by installing different modular units. Specifically, the new RMT system is composed of two standardized elements [16]:

- modular bed: a frame structure hosting two couples of helical racks. The modular beds, with a plant standard size on an EUR pallet, can be assembled together to form a beds line and extend the working volume along the main X axis;
- cross-tables hosted on the couples of helical racks of the beds line. Cross-tables can be workpiece-holder tables or tool-holder tables. Each cross-table has a standardized

driving autonomous subsystem to move each cross-table independently from the other cross-tables along the X axis of the beds line. An open kinematic architecture permits to obtain different types of machining, by hosting on the base cross-table operating devices and supplementary linear or rotational axes.

Figure 2 shows a bed module (a), a beds line (b) and the base cross-table (c). Cross-tables can be customized to change the processing technology and the typology of machining; furthermore, it is possible to change the working volume by varying the length of the beds line, i.e. the main axis X, by adding or removing beds modules. The kinematics of the new RMTs provides the flexibility modularity and scalability necessary to increase or reduce the number of processes performed in series or parallel along the beds line. Figure 3 shows a beds line hosting on the upper helical racks a milling spindle with extra Y and Z axes.

3 Analysis

In this paper the comparison of the environmental impact of the new RMT and the conventional machine tools is conservative since it is made considering for the new RMT only the function provided by the conventional machine tool, specifically the milling operations, without considering its reconfigurability characteristics.

The comparison considers the new Profacere® RMT with milling setup (Fig. 3) comprising one bed module and the Haas TM-3P milling machine tool (Fig. 4). In particular LCA studies two phases: the production of the machine tools and their utilization. Tables 1 and 2 report respectively the specifications of the bed modules, the base customizable cross-table, the customized operating milling cross-table and the motor units data; the RMT has a 3.5 kW spindle. The specifications of the conventional machine tools are provided by the producer website [17].

The weights of the materials that compose the machine tools have been identified. For the new RMT assembled by using one modular bed and one milling cross-table, the weights are: steel 1075 kg, Aluminium 2 kg and brass 1 kg. For the conventional milling machine tools, only the steel has been considered with a weight totalling 2813 kg.

Considering the manufacturing of the RMT and the conventional milling machine tool, only the material masses are taken into account. Masses are the inputs of the manufacturing process and the outputs are the RMT itself and the milling machine tool.

The analysis results provided on the basis of the masses concern some pollutants emitted in the atmosphere and some impact categories. The pollutants taken into account are the carbon dioxide, the carbon monoxide, and the sulphur dioxide; Table 3 reports a comparison between the RMT and the conventional milling machine. Considering the impact categories, the ones chosen are the resources use concerning fossils, water and climate change-fossil. These parameters are probably the most incisive, significant and relevant in terms of emissions and environmental impact. Table 4 compares the impacts for the manufacture phase.

The analysis of the utilization phase examined only the required energy. For the new RMT the power absorbed by the machine module has been obtained by considering the four motors (Nema 34) for the axes motion and the spindle; the power absorbed by

the axes is 1.6 kW, the spindle power is 3.5 kW; the total power absorbed is 5.1 kW. Assuming this hypothesis, the maximum annual absorption for the RMT, estimating a usage of 8 h per day for one year, is 12556 kWh.

Fig. 4. Haas TM-3P.

Table 1. RMT specification of bed module, base cross-table and operating milling table.

Configuration	Specifications	Unit
Bed module		
Columns	100 × 100 × 4 Length = 1750	mm
X-axis horizontal traverse	100 × 100 × 4 Length = 590	mm
Y-axis horizontal traverse	100 × 100 × 4 Length = 980	mm
Traverse X-axis cap	100 × 50 × 4 Length = 790	mm
Traverse Y-axis cap	100 × 50 × 4 Length = 1180	mm
Lower caps for bases	Metal plate 100 × 100 × 10	mm
Cap fixing inserts	4 × 50 × 20 Length = 50	mm
Base extension frame	1 × 50 × 50 × 4 Length = 980 2 × 50 × 50 × 4 Length = 49	mm
Pipes DN25	DN25 Length = 980	mm
Plate shelf	998 × 978 × l.5	mm
Electrified racks	7 poles	
Door panel	1113,8 × 1550 × 1.50	mm
Side panel	690 × 1550 × 1.50	mm

(*continued*)

Table 1. (*continued*)

Configuration	Specifications	Unit
Threading Flange	DN25 PN6 threading	
Base cross-table		
Side plates S235	805 × 200 × 8	mm
Profile	80 × 40 × 4 Length = 963	mm
Welded frame mm	2 × 80 × 40 × 963 and 2 × 80 × 40 × 470	mm
Engine adjustment profile	90 × 50 × 8 Length180	mm
Operating milling table		
Welded frame	2 × 80 × 40 × 963 and 2 × 80 × 40 × 470	mm
C bearing frame	2 × UPN C220 × 80 × 9 Length = 963 2 × UPN C220 × 80 × 9 Length = 470	mm
Y- axis gear	241 × 241 × 410 × 410	mm
Z-axis section bar	180 × 180 × 14	mm
Z-axis slides thickness	25 × 20.4 Length = 700	mm

Table 2. Axes motor unit specification.

Configuration	Specifications	Unit
Nema 34 engines	7	Nm
Driver AC	110–230 4.5–8	Vac A

Table 3. Pollutants emissions for manufacture phase (units kg).

Emissions	CO_2	CO	SO_2
New RMT	2722	22904	4691
Conventional milling machine	7059	59768	12242

Table 4. Comparison of impacts on resources for manufacture phase.

Resource Unit	Fossils kWh	Water m^3	Climate change kg
New RMT	25085	403	2979
Conventional milling machine	64780	1041	7725

As in the case of the production phase, only the estimated annual electricity consumption was chosen as input. The outcomes are the carbon dioxide and the resource

Table 5. Comparison of emissions for utilization phase

Emissions	CO_2 Kg	Resource fossils kWh
New RMT	25100	17310
Conventional milling machine	39956	27560

use concerning fossils. Similar evaluation was performed for the conventional milling machine tool whose total absorbed power is 6.85 kW. A similar analysis for the conventional milling machine provides a yearly consumption of 19988 kWh. Table 5 compares the new RMT and the conventional milling machine for the utilization phase.

4 Discussion and Conclusions

LCA makes evident that the new RMT is more sustainable, since the amount of emissions obtained for the conventional milling machine are higher for the manufacturing and utilization phases than that concerning the RMT.

In particular the manufacturing phase shows larger emissions and impacts, almost triple, for the conventional milling machine. During the utilization phase the differences are not so large due to the specific operations that is identical for both the systems, TMT and conventional milling machine tool.

Furthermore, in terms of sustainability, the new RMT is also better when a two modular beds line is employed, instead of only one modular bed line. Therefore, even increasing the working volume, the new RMT is more sustainable since the values of environmental impacts and pollutant emissions would still be lower than those of larger conventional milling machine. Moreover, the modularity of the new RMT enables to renew its elements, when defective, and thus extend the operative life of the systems.

Another aspect that makes reconfigurable machine tools and reconfigurable manufacturing systems more sustainable is the possibility to enable the reuse of the manufacturing resources and reorganize the supply chain to move production near the end users and consequently reduce the emissions and impacts due to the transport [18].

References

1. Blum, H., Legey, L.F.L.: The challenging economics of energy security: ensuring energy benefits in support to sustainable development. Energy Econ. **34**(6), 1982–1989 (2012)
2. United Nations. Kyoto Protocol to the United Nations Framework Convention on Climate Change (1989). https://unfccc.int/resource/docs/convkp/kpeng.pdf
3. Humlum, O.: The State of the Climate 2021, Report 51, The Global Warming Policy Foundation (2022)
4. UN Environment Programme. United Nations Conference on Sustainable Development. https://www.unep.org/pt-br/node/23750

5. Ueda, K., Takenaka, T., Váncza, J., Monostori, L.: Value creation and decision-making in sustainable society. CIRP Ann. **58**(2), 681–700 (2009)
6. Sutherland, J.W., et al.: The role of manufacturing in affecting the social dimension of sustainability. CIRP Ann. **65**(2), 689–712 (2016)
7. Hazra, Somnath, Bhukta, Anindya (eds.): Sustainable Development Goals. SDGS, Springer, Cham (2020). https://doi.org/10.1007/978-3-030-42488-6
8. Elkington, J.: Cannibals with Forks: The Triple Bottom Line of 21st Century Business. Oxford Capstone Publishing (1997)
9. Scientific Applications International Corporation (SAIC). Life cycle assessment: principles and practice, U.S. EPA, Washington, DC, EPA/600/R-06/060 (2006)
10. Dassisti, M., Semeraro, C., Chimenti, M.: Hybrid exergetic analysis-LCA approach and the industry 4.0 paradigm: assessing manufacturing sustainability in an Italian SME. Procedia Manuf. **33**, 655–662 (2019)
11. Bruzzone, A.A., D'Addona, D.M., Rosciano, I.: Thermoeconomic analysis of LEGO®-like reconfigurable machine tools. Procedia CIRP **88**(2020), 375–380 (2020)
12. https://www.openlca.org/
13. Rodriguez, C.: ELCD 3.2 in openLCA, GreenDelta GmbH (2016)
14. Goedkoop, M., Spriensma, R.: The Eco-indicator 99 a damage oriented method for life cycle impact assessment (2001)
15. Landers, R.G., Min, B.-K., Koren, Y.: Reconfigurable machine tools. CIRP Ann. **50**(1), 269–274 (2001)
16. Bruzzone, A.: Advanced reconfigurable machine tools for a new manufacturing business model. In: Andersen, Ann-Louise., et al. (eds.) CARV/MCPC -2021. LNME, pp. 47–54. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-90700-6_4
17. https://www.haascnc.com/index.html
18. Napoleone, A., Bruzzone, A.A., Andersen, A.-L., Brunø, T.D.: Fostering the reuse of manufacturing resources for resilient and sustainable supply chains. Sustainability **14**(10), 5890 (2022). https://doi.org/10.3390/su14105890

Comparison of Machine Learning Models for Predictive Maintenance Applications

Alessia Lazzaro[1](✉), Doriana Marilena D'Addona[2], and Massimo Merenda[3]

[1] University Mediterranea of Reggio Calabria, 89124 Reggio Calabria, Italy
lazzaroalessia18@gmail.com
[2] University of Naples Federico II, 80125 Napoli, Italy
dorianamarilena.daddona@unina.it
[3] Center for Digital Safety Security, Austrian Institute of Technology GMbH, 1210 Vienna, Austria
massimo.merenda@ait.ac.at

Abstract. In the field of industry 4.0, one of the sectors in which research is particularly active is the area of Predictive Maintenance(PdM), the purpose of which is to improve the industrial production process. This type of maintenance aims to predict a possible failure event, reduce non-production times and increase the quality of the processing result. The objective of this paper is to select the best Machine Learning models for a PdM application. In particular, such a model should allow making a prediction based on a real dataset, obtained by monitoring a turning process, with the aim of making the classification of the chip shape. The criteria used to choose the best model are accuracy and prediction speed (to reduce the inference time). Indeed it is crucial to spot any potential machine fault in the shortest time possible, in order to intervene before the machine fails. Hence, our goal is to choose the ML models with a lesser inference time while still maintaining high accuracy.

Keywords: Machine learning · Predictive maintenance · Classification learner app · Chip form classification

1 Introduction

The improvement of the production strategy is one of the objectives of the 4^{th} industrial revolution, in order to decrease the production costs and, at the same time, increase the quality of the final product, to survive in an increasingly competitive global market. These needs have led to the redefinition of the machinery maintenance strategy, which today can leverage a technological advancement that allows increasing the efficiency of the entire production process. Moreover, this technological advancement provides benefits in terms of safety, decreases time-to-market, and provides a better estimation of the number of stocks needed.

The maintenance strategies applied in the industrial field can be divided into 3 macro-categories: *Corrective Maintenance*, *Preventive Maintenance* (PM), and *Predictive Maintenance* (PdM).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 657–666, 2023.
https://doi.org/10.1007/978-3-031-16281-7_62

Corrective Maintenance is unplanned maintenance whose only goal is to bring the item back to a functioning state as soon as possible, either by repairing or replacing the failed item or by switching in a redundant item [1].

Preventive Maintenance (PM) is scheduled maintenance where periodic checks are scheduled in advance and the frequency of the activity can be chosen according to different factors. PM can lead to organizing production stops and can reduce the probability of failure but it cannot eliminate the occurrence of catastrophic failure.

Predictive Maintenance (PdM) is a technique that allows to optimize the use interval of the equipment while reducing the frequency of maintenance activities to a minimum through the real-time health monitoring of the assets to predict failures and replace the components just before their breakage. The monitoring activity provides a continuous flow of information thanks to which it is possible to find a potential anomaly, that is, to perform a prediction of the state of the in-service equipment. According to [2], PdM enables to lessen 25%–35% the maintenance costs, eliminate 70%–75% breakdowns, reduce 35%–45% breakdown time, and increase 25%–35% production.

Among the various approaches that can be distinguished in PdM, the Data-driven one (called the Machine Learning approach) is the most popular approach. To ensure the success of this PdM approach for a particular application it is necessary to make an accurate choice of Machine Learning (ML) models.

Therefore, the aim of this paper is to compare several ML techniques using different criteria in order to find out the ML model that better fits a practical example. In particular, we focus on the application of ML models to carry out the Predictive Maintenance of a turning machine, classifying the shape of the chip on the basis of the forces applied by the tools on the material. The criteria we use to choose the best model are accuracy and prediction speed (to reduce the inference time). Indeed it is crucial to spot any potential machine fault in the shortest time possible, in order to intervene before the machine fails. We recall that the inference time is one of the three components that contributes to the overall time to obtain the final prediction: the time interval for sending data from the sensor to the model, the inference time, and the time interval required to transmit and obtain the inference made.

To conclude this section, we have to point out that the vast majority of work on PdM applications focuses on the choice of ML models based on their probability of success in predicting a machine failure. Recently there is an emerging trend [3–5] aiming to take into account also the prediction process speed, with the aim to improve the efficiency of this process. However, to the best of our knowledge, in the field of chip form classification, our paper is the first to take into account the prediction speed in the choice of the best ML model.

This paper is organized as follows. In Sect. 2, we investigate the related literature. In Sect. 3 the maintenance problem is presented. In particular, in Subsect. 3.1 we describe the dataset used to train the models, in Subsect. 3.2 we explain the MATLAB application used in this paper to train the

ML models, and in Subsect. 3.3, we report some considerations on the results obtained. Finally, in Sect. 4, we draw our conclusion (Table 1).

Table 1. Abbreviations

PdM	Predictive Maintenance
PM	Preventive Maintenance
ML	Machine Learning
NN	Neural Network
ANN	Artificial Neural Network
SVM	Support Vector Machine
RA	Regression Analysis
BNN	Bayesian Neural Network
KNN	K-Nearest Neighbour
CNN	Convolutional Neural Network
RF	Random Forrest
CNC	Computer Numerical Control
PCA	Principal Component Analysis
SVR	Support Vector Regression

2 Related Work

In the scientific literature, there are several proposals concerning the application of ML models in specific fields of the manufacturing industry. In the following, we examine the proposals concerning the turning and the milling processes. The challenges faced in the literature are varied: from the prediction of cutting forces to the prediction of surface roughness, from the classification of the chip to the prediction of tool wear, etc.

In [7], to achieve the goal of classification for the chip shape in the turning process, an ANN is adopted, using three components of the cutting force as input. The same model is used by paper [6], which focuses on the problems of chip form classification and tangential cutting force prediction in turning operation. The final results showed an accuracy of 86.67% for classification chip form and accuracy of 91.13% for the main cutting force prediction. A Neural Network model is used by [8], with an emission signal analysis as input, to predict the chip form during the cutting process. The obtained accuracy is very high, around 80%. Chip form classification and favourable/unfavourable chip type prediction are the goals pursued by [9]. To archive these goals, the authors use feed-forward backpropagation neural networks reaching an accuracy greater than 90%. [10] uses an unsupervised neural network with real-time cutting force sensor signal spectrum as input, in order to control the chip formation during longitudinal turning. ANN is not only used for classification, as seen in the aforementioned examples but also for regression problems. In fact, [11] finds out that a Neural Network is more suitable for this task than a traditional linear regression model, since it shows a greater accuracy. In order to tackle the problem of predicting the specific cutting force, [13] uses an ANN feedforward. In particular, it was trained following an iterative process, which has a vector of 13 components as input, all the influencing parameters for this specific case.

The authors of the paper [12] show how neural network works better than the various regression models in predicting surface roughness. To predict the surface roughness in the milling process, the paper [14] makes a comparison between regression analysis (RA), SVM, and BNN (Bayesian neural network) and obtains a very similar prediction error values close to 8%. In [15] an ANN is used to monitor tool wear of a CNC machine, among different trained models since the result shows greater efficiency of the ANN model compared to the SVM and KNN models. The paper [16], in order to implement a tool breakage detection system, uses an SVR model with cutting forces in input. The obtained result is compared with the traditional regression approach and it is observed that the SVR model has slightly higher accuracy. However, the proposed model has to be trained through a relatively complicated tuning process of design parameters. In [17], a maintenance activity is used to monitor the progression of lateral wear of the tool, a CNN model and the accuracy obtained is 90%; while the authors of [18] make a comparison between ANNs, SVR, and RF algorithms. Although, the used ANN has one single hidden layer, which in this case is not enough. Indeed it is not suitable for the considered task, since it is known that ANN better performs with more hidden layers.

3 ML Models Comparison

In this paper, we aim to monitor the longitudinal turning process of carbon steel, with metal-coated inserts through the chip form classification. It is known that the turning process produces different chip shapes, which depend on the three components of the cutting force (F_f, F_c, F_p), represented in Fig. 1. Our goal is, therefore, to obtain a ML model that can autonomously predict the shape of the chip produced based on the cutting forces measured during the process.

The problem of chip formation has always been the focus of numerous studies and, with the evolution of production machinery, it has assumed ever greater importance. The ISO 3685 standard [19] defines three different chip shapes: *short spiral* (Fig. 2a), *short* (Fig. 2b), and *snarled* (Fig. 2c).

The problem of the formation of chips that are too long lies in the fact that in many cases it complicates the manufacturing process, as a chip that is too long can tangle around the tool. This can hinder the machining process, damage both the tool and the surface of the machined product (with a consequent decrease in the surface quality), and

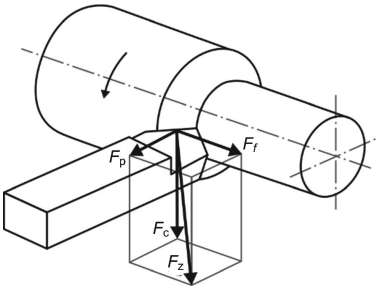


Fig. 1. Cutting forces generated during the turning process (Figure taken from [22])

(a) Short spiral (b) Short (c) Snarled

Fig. 2. Chip form

become a danger for operators since it leads to greater difficulties in evacuation from the work area.

Therefore, verifying the shape of the chip produced is of fundamental importance to understand if the process of turning is proceeding in the best way, otherwise, it is advisable to stop the operation, in order to avoid the issues mentioned above.

3.1 Data Acquisition

The dataset [23] consisting of twelve parameters and the respective target, is obtained by monitoring the components of the cutting force with force sensors. The cutting force component signals were digitized at 2500 Hz for 3 s providing a data sequence of 7500 points. The analysis of cutting force signal (CFS) specimens was carried out by achieving spectral estimation through a parametric method. From the CFS specimens (measurement vector), p features or predictor coefficients (feature vector), $a_1, \ldots, a_p$, characteristic of the spectrum model, are obtained through linear predictive analysis (LPA). Feature extraction was executed through the application of Durbin algorithm and the p value was chosen by examining the plot of the normalized rms prediction error vs. the order of the model.

The cutting parameters set to obtain the cutting forces, are the following:

- **Cutting speed** = 150, 250 m/min
- **Feed** = 0.08, 0.13, 0.20, 0.30 mm/rev
- **Depth of cut** = 1.0, 1.5, 2.0, 3.0 mm.

From the above process, it is obtained a dataset composed of 210 observations, characterized by 12 features and a label that can assume 3 integer values, divided as follow: 45 observations for class 0 which represents the *snarled* shape of the chip; 90 observations for class 1 which indicates the *short* shape, and 75 observations for class 2 which represents the *short spiral* shape.

Therefore, our aim is to train a multi-class supervised classification model which, starting from the value of the cutting forces, is able to infer the chip shape produced by the turning process.

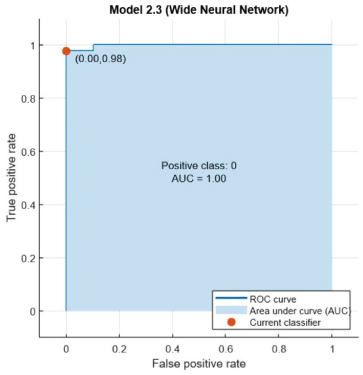
Table 2. MATLAB results.

Classifier	Classifier type	Accuracy (%)	Prediction speed (Object/Seconds)	Training time (Seconds)
Decision trees	Fine Tree	92.9	13000	2.8492
	Medium Tree	92.9	14000	0.5717
	Coarse Tree	77.6	11000	0.49467
Discriminant analysis	Linear Discriminant	74.8	10000	5.3121
	Quadratic Discriminant	99.0	7600	0.64689
Support vector machines	Linear SVM	73.8	6000	1.2843
	Quadratic SVM	98.1	5700	0.96552
	Cubic SVM	98.6	7700	0.67184
	Fine Gaussian SVM	98.6	6300	0.7483
	Medium Gaussian SVM	99.5	5200	0.64732
	Coarse Gaussian SVM	74.3	6300	0.79148
Nearest neighbors	Fine KNN	99.0	6800	0.56903
	Medium KNN	98.6	7300	0.45959
	Coarse KNN	51.9	7700	0.50991
	Cosine KNN	99.5	6500	0.49429
	Cubic KNN	97.6	5400	0.57902
	Weighted KNN	99.0	6600	0.49403
Naive Bayes	Gaussian Naïve Bayes	69.5	11000	1.1766
	Kernel Naïve Bayes	84.3	1700	1.9708
Ensemble	Boosted Trees	42.9	16000	0.65469
	Bagged Trees	98.1	2300	1.9497
	Subspace Discriminant	71.9	1400	1.8266
	Subspace KNN	99.5	760	2.8046
	RUS Boosted Trees	96.7	1800	2.0158
Neural network	Narrow NN	96.2	14000	0.78471
	Medium NN	97.1	15000	0.59822
	Wide NN	98.6	10000	0.70613
	Bilayered NN	96.2	12000	0.6635
	Trilayered NN	95.6	11000	0.83352

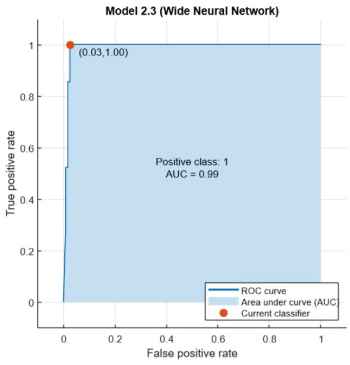
3.2 Models Training

To pursue our goal, we use *Classification Learner App*, included in the Statistics and Machine Learning Toolbox for MATLAB R2021, in order to select the best models. Indeed, this application can be used to train supervised and semi-supervised learning algorithms for binary and multi-class problems [20]. This application allowed us to train

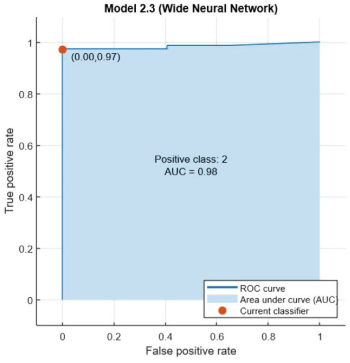
different models. Specifically, they are grouped as listed below: Decision Trees, Discriminant Analysis, Support Vector Machines, Nearest Neighbors, Naive Bayes, Ensemble and Neural Network.



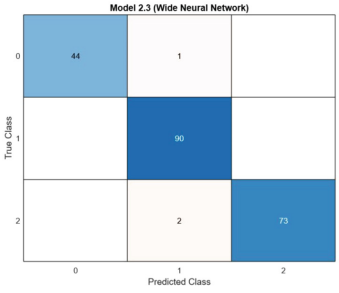
(a) ROC curve for class 0



(b) ROC curve for class 1



(c) ROC curve for class 2



(d) Confusion Matrix

Fig. 3. Performance of wide neural network: validation confusion matrix and validation ROC curve

Each group contains one or more models for a total of 29 models available in the R2021a version. We report the obtained results in Table 2, in which, for each model, we specify: Classifier family, Classifier type, Accuracy (%), Prediction speed (Objects/Seconds), and Training time (Seconds). The models were trained without using the PCA. Classification Learner App offers three possible choices for the model validation: *K-Fold Cross Validation* (K-FCV), *Holdout*, and *No Validation*. Among these, we choose to use the K-FCV method, setting the number of folds equal to 5. This method leads to better prediction accuracy, as it allows to avoid the overfitting issue. Once the application trains each model, it reports the details in the 'Summary' section. It is also possible to use the 'Optimizer' option which allows choosing the search range of the hyperparameters.

3.3 Results and Discussions

From the obtained results (reported in Table 2) we draw some considerations. Different models have high accuracy, above the 90%, but we can say that Neural Networks perform better than the other algorithms. Indeed, under the same accuracy, they need a shorter time to carry out the prediction and less time for the training phase. In addition, this type of algorithm is one of the most used in the field of Predictive Maintenance, via the ML approach, as also reported in Sect. 2. This is due to the fact that Neural Networks show high robustness and they return a very high accuracy in different scenarios, such as predicting chip shape, predicting specific cutting force, surface roughness and tool life, etc. Indeed, from the study conducted in [21] regarding the algorithms used for the prediction in the manufacturing sector, it appears that 28% of the problem of PdM in the manufacturing industry, the highest value, is solved using an ANN model. Therefore, focusing on NN, 5 types were trained by the MATLAB Toolbox: Narrow NN; Medium NN; Wide NN; Bilayer NN, and Trilayer NN.

For all models, MATLAB uses: *ReLU* as the activation function for all the fully connected layers, excluding the final fully connected layer where the activation function is always softmax; 1000 as the maximum number of training iterations; *Regularization strength (Lambda)* equal to 0, and activated data standardization.

Narrow NN has a single fully connected hidden layer made up of 10 neurons. *Medium NN* has a single fully connected hidden layer made up of 25 neurons. *Wide NN* has a single fully connected hidden layer made up of 100 neurons. *Bilayer NN* has two fully connected hidden layers, each with 10 neurons.

Trilayer NN has three fully connected hidden layers, each with 10 neurons. Classification Learner App allows visualizing the performances of the trained ML models. In the following, we report the Validation Confusion Matrix and the Validation ROC Curve in Fig. 3 for the Wide Neural Network.

4 Conclusion

In this paper, we compared several ML techniques by using two criteria (accuracy and prediction speed) in order to find out the ML model that better fits a practical application. In particular, we focused on the application of ML models to carry out the Predictive Maintenance of a turning machine, classifying the shape of the chip based on the forces applied by the tools on the material. Based on a real dataset, the comparison between possible classification models was made using Classification Learner App Toolbox. The results indicate that Neural Networks represent the best compromise between predictive speed and accuracy, from 97% to 99.5%.

An interesting goal for future work is to implement the best ML models, found out in this paper, on a microcontroller [24], in order to simulate a real-life application in the industrial field [25–27]. In a real-life application, one of the main concerns is that the time between data sensing (performed by the sensors positioned on the machinery) and the time in which the inference is received can be too high for some applications. Therefore, in this paper, we focused on the reduction of the inference time (i.e., choosing the ML models with the lesser inference time and still maintaining a high accuracy).

References

1. Rausand, M., Hoyland, A.: System Reliability Theory: Models, Statistical Methods, and Applications, vol. 396. Wiley (2003)
2. Van, T., Chan, I., Parthasarathi, S., Lim, C., Chua, Y.: IoT and machine learning enable predictive maintenance for manufacturing systems: a use-case of laser welding machine implementation. Available At SSRN 4073901 (2022)
3. Ouadah, A., Zemmouchi-Ghomari, L., Salhi, N.: Selecting an appropriate super-vised machine learning algorithm for predictive maintenance. Int. J. Adv. Manuf. Technol. **119**, 4277–4301 (2022)
4. Lin, S.: Application of machine learning to a medium Gaussian support vector machine in the diagnosis of motor bearing faults. Electronics **10**, 2266 (2021)
5. Pan, Z., Meng, Z., Chen, Z., Gao, W., Shi, Y.: A two-stage method based on extreme learning machine for predicting the remaining useful life of rolling-element bearings. Mech. Syst. Signal Process. **144**, 106899 (2020)
6. Suksawat, B.: Chip form classification and main cutting force prediction of cast nylon in turning operation using artificial neural network. In: ICCAS 2010. IEEE (2010)
7. Viharos, Z.J., Markos, S., Szekeres, C.: ANN-based chip-form classification in turning. In: Proceedings of the XVII. IMEKO World Congress–Metrology in the 3rd Millennium (2003)
8. Sukvittayawong, S., InaSaki, I.: Identification of chip form in turning process. JSME Int. J. Ser. 3, Vib. Control Eng. Eng. Ind. **34**(4), 553–560 (1991)
9. Segreto, T., Simeone, A., Teti, R.: Chip form classification in carbon steel turning through cutting force measurement and principal component analysis. Procedia CIRP **2**, 49–54 (2012)
10. Keshari, A., D'Addona, D., Teti, R.: Subtraction of inconsistence sensor data to improve the chip form classification and monitoring efficiency. In: 6th CIRP International Conference on Intelligent Computation in Manufacturing Engineering-CIRP ICME, vol. 8 (2008)
11. Hanief, M., Wani, M.F., Charoo, M.S.: Modeling and prediction of cutting forces during the turning of red brass (C23000) using ANN and regression analysis. Eng. Sci. Technol. Int. J. **20**(3), 1220–1226 (2017)
12. Nalbant, M., Gokkaya, H., Toktaş, İ.: Comparison of regression and artificial neural network models for surface roughness prediction with the cutting parameters in CNC turning. Model. Simul. Eng. **2007**, 14 (2007). Article No. 92717. https://doi.org/10.1155/2007/92717
13. Wenkler, E., Arnold, F., Hänel, A., Nestler, A., Brosius, A.: Intelligent characteristic value determination for cutting processes based on machine learning. Procedia CIRP **79**, 9–14 (2018)
14. Lela, B., Bajić, D., Jozić, S.: Regression analysis, support vector machines, and Bayesian neural network approaches to modeling surface roughness in face milling. Int. J. Adv. Manuf. Technol. **42**, 1082–1088 (2008)
15. Hesser, D.F., Markert, B.: Tool wear monitoring of a retrofitted CNC milling machine using artificial neural networks. Manuf. Lett. **19**, 1–4 (2019)
16. Cho, S., Asfour, S., Onar, A., Kaundinya, N.: Tool breakage detection using support vector machine learning in a milling process. Int. J. Mach. Tools Manuf. **45**(3), 241–249 (2005)
17. Gouarir, A., Martínez-Arellano, G., Terrazas, G., Benardos, P., Ratchev, S.: In-process tool wear prediction system based on machine learning techniques and force analysis. Procedia CIRP **77**, 501–504 (2018). 8th CIRP Conference on High Performance Cutting (HPC 2018)
18. Wu, D., et al.: A comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using random forests. J. Manuf. Sci. Eng. **139**(7) (2017)
19. Technical Committee ISO/TC 29: Tool-life testing with single-point turning tools, 2nd edn., November 1993. Reference number: ISO 3685:1993(E)

20. https://www.mathworks.com/help/stats/classification-learner-app.html. Accessed 11 Apr 2022
21. Du Preez, A., Oosthuizen, G.A.: Machine learning in cutting processes as enabler for smart sustainable manufacturing. Procedia Manuf. **33**, 810–817 (2019)
22. Zhu, K.: Modeling of the machining process. In: Smart Machining Systems. Springer Series in Advanced Manufacturing. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-87878-8_2
23. D'Addona, D., Teti, R.: Kohonen maps for chip form classification in turning. In: 3rd Virtual Conference on Intelligent Production, Machines and Systems - IPROMS2007, 2–14 July, pp. 630–635 (2007). ISBN 978-14200-7347-8
24. Merenda, M., Porcaro, C., Iero, D.: Edge machine learning for AI-enabled IoT devices: a review. Sensors **20**, 2533 (2020). https://doi.org/10.3390/s20092533
25. Merenda, M., Astrologo, M., Laurendi, D., Romeo, V., Della Corte, F.G.: A novel fitness tracker using edge machine learning. In: 2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON), pp. 212–215 (2020). https://doi.org/10.1109/MELECON48756.2020.9140602
26. Merenda, M., Porcaro, C., Della Corte, F.G.: LED junction temperature prediction using machine learning techniques. In: 2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON), pp. 207–211 (2020). https://doi.org/10.1109/MELECON48756.2020.9140539
27. Merenda, M., Cimino, G., Carotenuto, R., Corte, F.G.D., Iero, D.: Device-free hand gesture recognition exploiting Machine Learning applied to RFID. In: 2021 6th International Conference on Smart and Sustainable Technologies (SpliTech), pp. 1–5 (2021). https://doi.org/10.23919/SpliTech52315.2021.9566385

Exploring Manufacturing System Development and the Use of Platforms to Reduce Time-to-Market

Morten Skogstad Nielsen[1], Ann-Louise Andersen[1(✉)], Thomas Ditlev Brunoe[1], Khaled Medini[2], and Kjeld Nielsen[1]

[1] Department of Materials and Production, Aalborg University, Fibigerstraede 16, 9220 Aalborg East, Denmark
ala@mp.aau.dk

[2] Mines Saint-Etienne, Univ Clermont, INOP Clermont Auvergne, CNRS, UMR 6158 LIMOS, 42023 Saint-Etienne, France

Abstract. In the development and launch of new products to the market, manufacturing companies are continuously pressured to reduce time and cost to sustain competitiveness. In this regard, the effective and efficient development and ramp-up of manufacturing systems is a critical factor. However, while agile methods for increasing the speed of product design and development have received wide attention in both research and practice, manufacturing system development has in contrary been less explored and generally regarded as a less systematic or as an ad-hoc task. Therefore, this paper presents a multiple case-study from a large Danish manufacturing company with the aim of exploring challenges in three different manufacturing system development projects. The within-case and cross-case findings indicate that manufacturing system development indeed is a critical task, where many challenges regarding product to manufacturing system interface and internal interfaces in the manufacturing system architecture exists. Furthermore, findings indicate that lack of interface management leads to the redesign of new manufacturing systems, rather than reuse and reconfiguration. Thus, the case findings suggest a potential in increasing the use and reuse of existing manufacturing systems and explicitly designing platforms for reducing development cost and time.

Keywords: Manufacturing system development · Platform · Case-study · Changeable manufacturing · Manufacturing development

1 Introduction

In most manufacturing companies, pressure for cost and time reduction during new product development and launch are inevitable conditions that must be managed in order to reduce time-to-market and sustain competitive advantage. Most often, such new product design, development, or realization processes and projects are conducted either as a staged-gate approach or as a largely iterative, ad-hoc, or informal approach [1]. Methods for increasing the speed and effectiveness of the product design and development

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 667–676, 2023.
https://doi.org/10.1007/978-3-031-16281-7_63

task have received wide attention in both research and practice, e.g., integrated product development [2], concurrent engineering [3], generic product development process [4], modular and platform-based approached to product design [5], etc. However, the manufacturing system development is usually merely regarded as a less systematic last step in the product development process, which is conducted after a product and business case have been specified [1]. Moreover, compared to product development research, research and theories on the manufacturing system development process are limited. While Bellgran and Säfsten [1] thoroughly examine both the process and project perspective of developing new systems, research focusing on the development of manufacturing systems is scarce. Various paradigms for manufacturing systems exist, e.g., lean, flexible, reconfigurable, intelligent, personalized, human-centric, adaptive cognitive, etc. [6–8], but a general tendency in previous research is that the product development process in its entirety has received a lot of focus, while the manufacturing system development process has been regarded mostly in terms of partial theories devoted to manufacturing sub-problems or overall philosophies for manufacturing [1, 12]. Various methods for integrating the development of products and manufacturing systems in a modular and platform-based way have been proposed [9–11], however, there are currently no empirical investigations of how modularity and platforms can support challenges and deficits in manufacturing system development and how to improve this in manufacturing companies. Consequently, there is also a lack of empirically-based knowledge of how the manufacturing system design and development task is conducted in practice and which challenges should be addressed to increase the effectiveness and speed of this task, thereby reducing time-to-market for new products. Therefore, this paper explores different cases of manufacturing system development in a Danish manufacturing company with the aim of identifying challenges in the development of new manufacturing systems. The remainder of the paper is structured as follows: Sect. 2 presents the research method, while Sect. 3 presents the findings of the multiple case-study and the comparative cross-case findings. Section 4 discusses the theoretical and practical implications of findings, and Sect. 5 briefly summarizes the contribution of the paper and outlines future viable research directions.

2 Research Method

To address the research objective stated above, a multiple-case study methodology is applied. Case study research is applied, as this is a well proven method to build theory in an industrial context [13]. Furthermore, case study research is particularly appropriate when analyzing and building theory based on few observations, where many variables are interacting in a complex system [14]. In this research, previously published literature is scarce as indicated in the previous section, which calls for exploratory case research. Thus, in this paper, the unit of analysis is the manufacturing system development project, where multiple cases of these are studied in order to allow for cross-case comparisons and more general implications.

2.1 Case Company and Selection of Cases

The case company is a market-leading large enterprise with headquarters located in Denmark and numerous manufacturing plants and sales offices worldwide. Within this company, product and manufacturing systems development is traditionally handled largely as separate projects with individual scope, key performance indicators, budgets, time-plans and objectives. Moreover, manufacturing system development is often conducted when a product and market demand have been specified to large extent, resulting in highly dedicated solutions for each new products or variants. However, in the latest years, the company has experienced unsatisfactory performance of development projects. Thus, in order to increase competitiveness, the company focuses not only reducing development time and cost, but also on exploring the potentials for increasing reuse in manufacturing through platforms, modularity, and increased reconfigurability. Three development projects conducted in the case company within the last two years were selected for this research (denoted as Case A, Case B and Case C). These three specific cases were chosen, as they have different overall objectives and scope, thereby representing the variety of development projects withing the company. Moreover, each of the cases reached unsatisfactory performance, thus, being ideal for studying challenges inherent in manufacturing system development and in relation to new product development.

2.2 Data Collection and Analysis

In each of the three cases, multiple sources of evidence were used to study challenges occurring in the development process and project. Primary data consisted of a number of semi-structured interviews conducted with relevant employees e.g., process engineers, manufacturing directors, project managers in each project. A total of 10 semi-structured interviews were conducted. In addition, archival records such as internal company presentations, project documentation, and performance data were collected. For each of the three cases, the collected qualitative and quantitative data was analyzed with the purpose of uncovering critical challenges in development, followed by cross-case analysis with the aim of studying general challenges and potentials in manufacturing development to increase performance. During the data gathering phase, one of the authors have been working in the case company, by performing the interviews, gathering historical data in

Table 1. Case protocol used in studying the three cases (Case A, Case B, Case C)

Characteristic of case	Information collected in case
Scope of development	Type of product, Newness of product, Time of project, Objectives
Project structure and type	Participants
Performance	Performance of project in terms of time plan and budgets, Challenges experienced by the participants
Type of manufacturing development	New technology, scope of products covered, newness of equipment etc.

terms of documents, evaluations and quantitative data on development projects. During this phase, the authors made no interventions to ensure validity of the findings. After the data gathering, the cases were analyzed individually and comparatively to establish cross case findings. In Table 1, the characteristics of each studied case is outlined along the information collected in the cases.

3 Case Findings

In this section, the findings from each individual case are initially reported, where Table 2 outlines the performance overview of each studied development project. Evidently, each of the three studied cases experienced unsatisfactory performance, and the following reported case findings outline the main challenges identified as causes for this. In the last part of this section, these findings from the individual cases are cross-examined in order to compare and contrast findings towards more generalizable results.

Table 2. Overview of the performance of each case compared to budgets.

Characteristic	Case A	Case B	Case C
Project cost (manufacturing development)	+13%	+78%	+170%
Investment (manufacturing equipment)	+4.6%	−12%	0%
Time-to-market	+29% (+158% full operational)	+146%	+42%
Development hours	>12.000 h	>23.000 h	>15.000 h

3.1 Case A

The overall objectives of case A were to develop a new version of an already existing product, in order to comply with future market requirements. For simplicity, this product is made of two major parts: part 1 (Electronic) and part 2 (Mechanical). In order to produce this product, the following manufacturing lines are required: a final assembly line, two assembly lines producing part 1 or part 2 and multiple component lines producing components for either part 1 or part 2. Figure 1 illustrates the relation between the newness of both the product and the manufacturing systems in case A.

In case A, the major challenges occurred in two different areas, firstly in the assembly and test of part 1 and secondly in the final assembly of the finished product. In part 1, the new product design exceeded the capabilities of a single piece of manufacturing equipment resulting in modification of the component manufacturing line. This misalignment between product design and manufacturing capabilities was a direct consequence of not having the capability of the existing manufacturing equipment documented and therefore not available to the product designers. As part 1 is an electronic part, the manufacturing of this is done on a highly automated/high volume manufacturing line with some

build in flexibility. Therefore, this modification, described with "1" in Fig. 1, was not desirable and lead to cost over-runs and time delays. When designing the new assembly line for part 1, described with "2" in Fig. 1, the main problem was the delivery of a fixture for the testing equipment. This fixture had a long lead-time and could only be ordered when the product-to-manufacturing interface was locked. This product-to-manufacturing interface experienced late changes because the product architecture was changed due to limitations in the component manufacturing line for part 1. The problem in final assembly, described with "5" in Fig. 1, was not represented in the performance of this project, described in Table 2, as these were discovered early in the project and therefore accounted for in the budgets and time schedules. However, designing a whole new final assembly line was only found necessary because the existing line was designed specifically for the previous product and with no considerations for potential future versions. This was done to gain the lowest possible unit-cost for the existing version. When designing this new final assembly line, no formal requirement for reuse of previous solutions where made and information was only passed on through tacit knowledge by the designers and engineers that was part of both development projects. This challenge was a result of not utilizing information in a previous project and thereby not preparing the manufacturing for potential future projects. This challenge was passed on to future projects as these was again not included. This problem caused the 158%-time overrun described in Table 2, before the new final assembly line was fully operational.

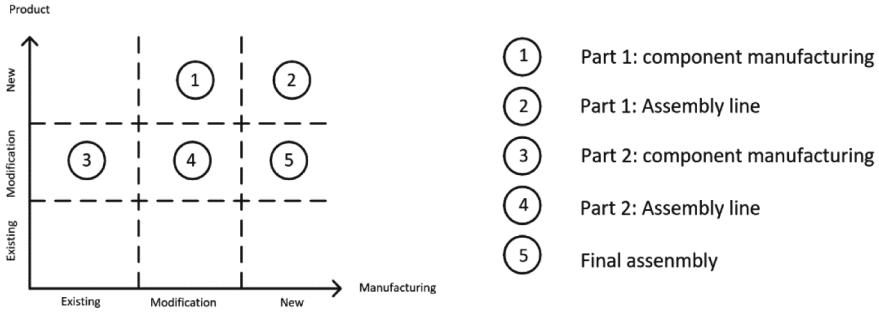


Fig. 1. Product developments impact on manufacturing in case A.

3.2 Case B

The overall objectives of case B were to develop a new product (based on existing technology) in order to penetrate a new market. Because physical size is an important driver in the specific market, the product architecture of this new product was highly integrated. However, for simplicity in this context, the product is said to consist of three major parts (two mechanical and one electronic). In order to produce this new product, the following manufacturing lines were required: a final assembly line and multiple component lines for each of the three major parts. In Fig. 2, the relations between the newness of the product and the manufacturing systems in this project is outlined.

In case B multiple challenges occurred. Here again, two major challenges occurred because the new product design exceeded the capabilities of the exiting manufacturing equipment. In one situation, a component was designed with tolerances smaller than what the existing manufacturing equipment was capable of and at the edge of what the current technology could provide. As in Case A, this misalignment between product design and manufacturing capabilities was a direct consequence of not having the capability of the existing manufacturing equipment documented and therefore not available to the product designers. This resulted in the design of a new component manufacturing line described with "1" in Fig. 2. The manufacturing process required for the assembly of two components was beyond the capabilities of the existing manufacturing equipment. Furthermore, the late changes to the product design caused by the problems in the component manufacturing caused extra delays because the final testing was not able to be designed before the final product design was locked. These two challenges accounted for a substantial part of the budget overruns on development cost and time-to-market described in Table 2. Similar to case A, late change is the product design resulted in delays. However, unlike case A this was caused both by changes in product-to-manufacturing interfaces, but also in change of requirements for the manufacturing system which led to late changes in the manufacturing setup.

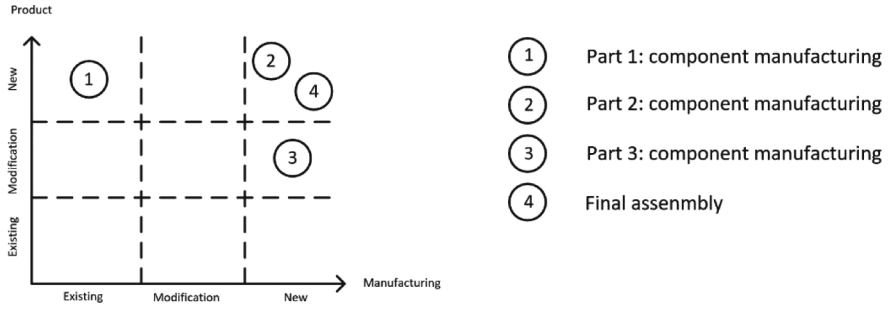


Fig. 2. Product developments impact on manufacturing in Case B.

3.3 Case C

The overall objective of case C was to develop three modular components which could be sold as a finished product or could be further integrated into a range of existing products. For simplicity these products are made of two major parts: part 1 (Electronic) and part 2 (Casing). In order to produce these new components, the following manufacturing lines are required: one assembly line and one component manufacturing line. Figure 3 outlines the relation between product development and manufacturing systems in case C.

In case C, one major challenge occurred which cascaded into several other challenges. In the newly designed assembly line, the testing was very dependent on the physical size of the product, meaning that small alteration to the product would require changes to the testing process. This testing process was historically a bottleneck for manufacturing

design (also in case A & B) and therefore the testing equipment for the component assembly described with "2" in Fig. 3 was designed to be more flexible. This was done by creating a fixed interface between the test-fixture on manufacturing tool level (defining the product-to-manufacturing interface) and the remaining test equipment. However, this led to challenges in defining the future capabilities in terms of the physical size that the test equipment would be able to handle and to standardize test software to be used across the entire product scope and also be utilized on other products beyond the scope of this case. This challenge of developing a modular test architecture led to the significant cost overrun and time delay displayed in Table 2.

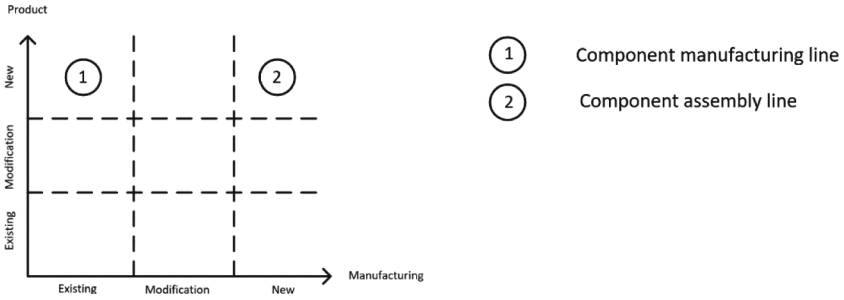


Fig. 3. Product developments impact on manufacturing in Case C.

3.4 Cross-Case Findings

Each of the cases reported above indicates several challenges in the development of new manufacturing systems, and in particular in relation to the product development complexity. In order to conclude whether any of these challenges found in the three cases can be perceived as reoccurring general problems, categorization and comparison is needed. Thus, in Fig. 4, the case challenges are categorized and compared in regards to three parameters; the factory structuring level on which the identified challenge was related to (i.e., line, cell, equipment and tool level), the type of manufacturing being developed (i.e., final assembly, sub-assembly, or component manufacturing), and whether causing cost, time or cost and time overruns.

In Fig. 4, it is evident that the challenges in all three cases mainly occur on the lower manufacturing levels such as on equipment and tool level. Only in case A, one problem occurred on a higher level because the existing final assembly line was too integrated. Moreover, all problems found across the three cases occur when developing the manufacturing of new components and the subassemblies of these. The challenges found within these two groups are all product-to-manufacturing interface and manufacturing capabilities related. Finally, it is evident that all of the challenges found within all cases caused either time or time and cost overrun, meaning that none of the challenges found only caused cost overruns.

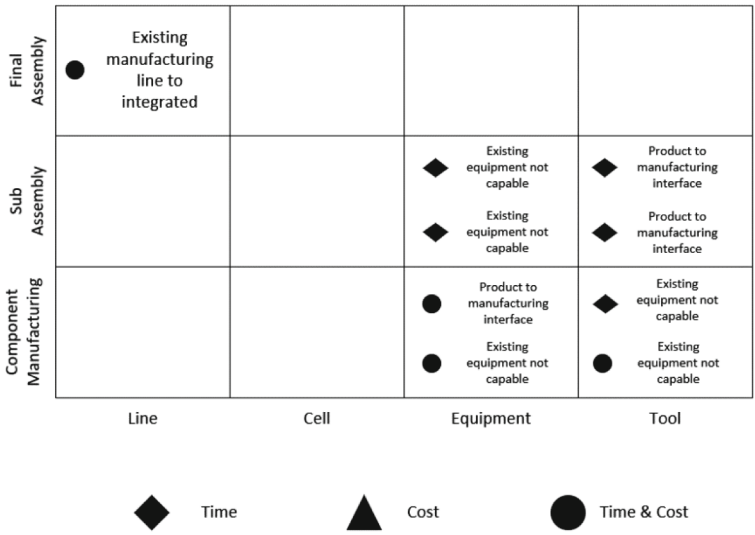


Fig. 4. Case challenges categorized and compared.

4 Discussion

The findings reported in this paper indicate critical challenges that a manufacturing company is facing when designing and developing manufacturing systems in order to introduce new product offerings. The main challenges were related mainly to the architecture of existing equipment and tools being too integrated and dedicated, thus hindering reuse in the new development efforts. Moreover, the product to manufacturing system interface often caused issues, in particular on equipment and tool level. While these findings appear generalizable across the three cases studied in the particular case company, further empirical evidence is needed to confirm the generalizability of these findings beyond the given company context. While previous research in this regard is scarce, Bruch and Bellgran [9] identified information management as a key enabler of an improved production system design process. In this research, different types of information and different situations where information was either not shared or not available was found. Firstly, information sharing between the product and manufacturing domains was a cause of several challenges. This problem was perceived to be on an operational level where product and manufacturing system designers did not share all relevant design information during the development phase. This lack of information sharing on an operational level caused delays and cost overruns, as changes in the product domain were not aligned with manufacturing. Secondly, information sharing on a tactical level was an issue, where reusing information from previous development project and the capabilities of the current manufacturing setup was not available. This lack of information caused product designers to design products with no regards to the capabilities of the current manufacturing setup, thereby not utilizing existing product to manufacturing interfaces and manufacturing capabilities. Finally, information sharing on a strategic level where

all the previous information together with future plans are used when developing manufacturing systems in order to mitigate the problems identified in this research. Similar to this, several studies report problems and challenges adjacent to the findings of this study [15–17], however, these are mostly explicitly in relation to the ramp-up period of new manufacturing systems, while this study focused on the entire process of developing new manufacturing systems.

5 Conclusions

This paper presented a multiple case-study from a large Danish manufacturing company with the aim of exploring challenges in three different manufacturing system development projects. Based on the findings in this research, it is evident that information regarding interfaces and manufacturing capability are important. These interfaces cover both i) the product to manufacturing system interface, which have been found to cause challenges is all cases in this research, as well as ii) the internal interfaces in the manufacturing system architecture, where lack of interface management lead to the redesign of new manufacturing systems. Also, information regarding manufacturing capability is found to be important as these information sets the boundaries for product development to comply with in order to reuse existing manufacturing setup. Challenges with information management that have been identified within this research have all been identified as causes of both cost overruns and time delays in the development of manufacturing systems. It is therefore recommended that future research focus on how to generate, maintain and utilize these types of information that has been identified in this research in order to reduce cost and time-to-market in future manufacturing system development projects. Thus, the case findings suggest a potential in increasing the use and reuse of existing manufacturing systems and equipment and explicitly designing platforms for reducing development cost and time. Future research should accordingly focus on: i) investigating the impact of product modularization and strategic partitioning on the manufacturing, ii) co-development of product and manufacturing and how to foster reuse in development projects, and iii) best-practice, formal rules, and guidelines for reusing previous manufacturing solutions as well as designing solutions with future plans in mind.

References

1. Bellgran, M., Säfsten, K.: Production Development: Design and Operation of Production Systems. Springer, London (2009). https://doi.org/10.1007/978-1-84882-495-9
2. Gerwin, D., Barrowman, N.J.: An evaluation of research on integrated product development. Manage. Sci. **48**(7), 938–953 (2002)
3. Prasad, B.: Concurrent Engineering Fundamentals, vol. 1. Prentice Hall PTR, Hoboken (1996)
4. Ulrich, K.T., Eppinger, S.D.: Product Design and Development. McGraw-Hill, New York (2004)
5. Ericsson, A., Erixon, G.: Controlling Design Variants: Modular Product Platforms. SME (1999)
6. Koren, Y., Gu, X., Badurdeen, F., et al.: Sustainable living factories for next generation manufacturing. Procedia Manuf. **21**, 26–36 (2018)

7. Gu, X., Koren, Y.: Mass-individualisation–the twenty first century manufacturing paradigm. Int. J. Prod. Res., 1–16 (2022)
8. ElMaraghy, H., Monostori, L., Schuh, G., et al.: Evolution and future of manufacturing systems. CIRP Ann. **70**(2), 635–658 (2021)
9. Bruch, J., Bellgran, M.: Integrated portfolio planning of products and production systems. J. Manuf. Technol. Manag. **25**(2), 155–174 (2014)
10. Michaelis, M.T., Johannesson, H.: From dedicated to platform-based co-development of products and manufacturing systems. In: ElMaraghy, H.A. (ed.) Enabling Manufacturing Competitiveness and Economic Sustainability, pp. 196–202. Springer, New York (2011). https://doi.org/10.1007/978-3-642-23860-4_32
11. ElMaraghy, H., Abbas, M.: Products-manufacturing systems co-platforming. CIRP Ann. Manuf. Technol. **64**(1), 407–410 (2015)
12. Michaelis, M.T.: Co-development of products and manufacturing systems using integrated platform models (2013)
13. Voss, C., Tsikriktsis, N., Frohlich, M.: Case research in operations management. Int. J. Oper. Prod. Manag. **22**(2), 195–219 (2002)
14. Yin, R.: Case Study Research: Design and Methods, 5th edn. SAGE Publications, Inc., Thousand Oaks (2014)
15. Javadi, S., Bruch, J.: Start of production in low-volume manufacturing industries: disturbances and solutions. In: Umeda, S., Nakano, M., Mizuyama, H., Hibino, H., Kiritsis, D., von Cieminski, G. (eds.) APMS 2015. IAICT, vol. 459, pp. 475–483. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22756-6_58
16. Surbier, L., Alpan, G., Blanco, E.: Identification of problem types during production ramp-up. In: International Conference on Industrial Engineering and Systems Management (2009)
17. Andersen, A.-L., Bejlegaard, M., Brunoe, T.D., Nielsen, K.: Investigating the impact of product volume and variety on production ramp-up. In: Bellemare, J., Carrier, S., Nielsen, K., Piller, F.T. (eds.) Managing Complexity. SPBE, pp. 421–434. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-29058-4_33

Sensor of Things

Automated Tonic-Clonic Seizure Detection Using Random Forests and Spectral Analysis on Electroencephalography Data

Craig Stewart[1(✉)], Wai Keung Fung[2], Nazila Fough[1], and Radhakrishna Prabhu[1]

[1] School of Engineering, Robert Gordon University, Aberdeen, UK
C.stewart35@rgu.ac.uk

[2] Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff, UK

Abstract. Artificial intelligence (AI) has a potential for impact in the diagnosis of neurological conditions, the academic consensus generally has a positive outlook regarding how AI can improve the care of stroke victims and those who suffer from neuro-degenerative conditions such as dementia. When combined with Internet of Things technology, this could facilitate a new paradigm for epilepsy treatment. These technologies have applications in improving the welfare of epileptics, epilepsy being a common neurological condition that can result in premature death without a quick response. As such it is important for the system to avoid false negatives. This investigation focused on how machine learning algorithms can be utilised to identify these events through Electroencephalography (EEG) data. The UCI/Bonn dataset, a classic benchmark for automated epilepsy detection systems was identified and utilised. This investigation focused on the random forest algorithm. Given that EEG neurological data represents time series data and machine learning excels at this task, automation could be achievable via a wearable device. From there, Fast Fourier Transforms (FFT) were applied to identify if spectral features of EEG signals would aid identification of seizures. This method achieved an accuracy of 99%, precision of 98% and a recall of 100% in 12.2 ms time to classify and one second of EEG data. These results show that random forests combined with FFT are a viable technique for attaining high recall when detecting grand mal epileptic seizures in short periods of time. CHB-MIT dataset was utilized for parity also showing good performance.

Keywords: Epilepsy · Seizure detection · Random forest · Fast Fourier Transform (FFT) · Artificial intelligence (AI) · IOT

1 Introduction

Epilepsy is a group of neurological conditions that result in recurrent seizures of various magnitudes. This condition displays heterogeneity in outcome, this is caused by sheer variation of source within the brain meaning that they can result in anything from relatively minor behavioural infringements to life threatening arrests. The cause usually is a form of neurological damage that can lead to either increased or reduced sensitivity

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 679–688, 2023.
https://doi.org/10.1007/978-3-031-16281-7_64

to stimulus triggering a chain reaction that spreads throughout an area of the brain as discussed in [1] by Thjis *et al*. The size of the area in the brain is related to the magnitude of the seizures as seen also in [1]. This family of conditions is relatively common throughout society with the World Health Organisation (WHO) estimating that 70 million people globally have some form as discussed in [2]. By in large, epileptic conditions are treatable with conventional medications, reducing the magnitude of the seizures or stopping them completely. However, there is a significant minority that will endure the condition throughout life experiencing a diminished quality of life with the treatment itself inducing side-effects as well as social barriers being imposed [1].

Epileptic seizures can be hazardous in of themselves; however, injuries can also be sustained from the uncontrolled nature of collapse, serious head injuries as a result can be sustained that prove fatal. Care is also hampered by the unpredictability of seizures with forewarning being limited often to the moments leading up to an event, often known as an "aura", a grace period that allows for preparation. The seizures themselves, in nature, can be unpredictable, individuals prone to tonic-clonic seizures are not excluded from experiencing an entire spectrum of seizures from absence to full tonic-clonic arrests. Thus, there is vested research interest into technological solutions capable of predicting when a seizure is likely to occur or otherwise provide prompt warning to carers as to if one has occurred. There are also applications within the healthcare profession where automatic epileptic event detection can speed up identifying of activity allowing for quicker diagnosis. Figure 1 illustrates the diversity epileptic conditions.

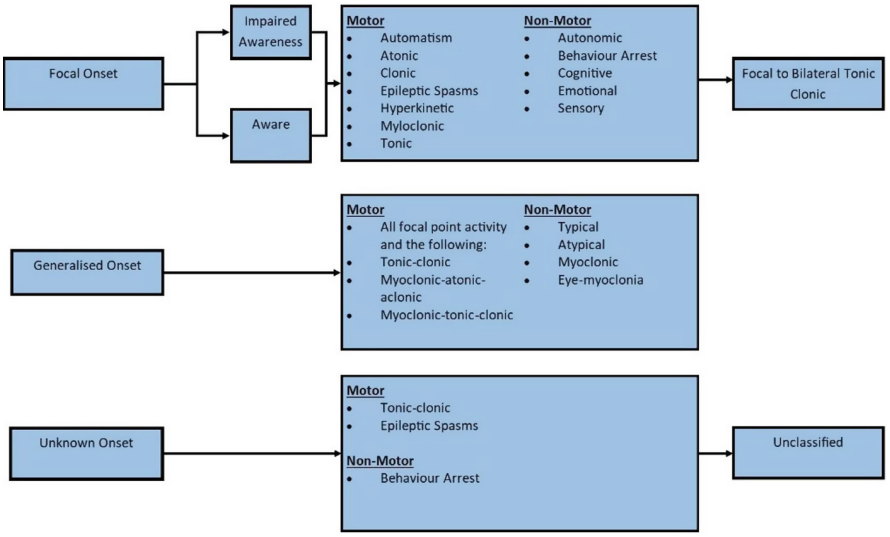


Fig. 1. The International League Against Epilepsy framework for the classification of epileptic seizures [3]

There are numerous signal types that can either implicitly or explicitly suggest that a seizure event has occurred, some academics elect to follow implicit, indirect signals related to motion and temperature changes related to possible tonic-clonic activity, the

embrace2 watch [4] takes such an approach whilst others work with explicit, direct signals from the brain itself utilising Magnetic Resonance Imaging (MRI) and Electroencephalography (EEG) [1]. Prediction and medication are typically done utilising EEG, this involves several electrodes being placed directly on to an individual's scalp to monitor the electrical activity occurring. Identification is made possible through specific signal traces that correspond with an event occurring, helping to segregate signals that correspond with normal and abnormal functions. Seizures have four types defined by their amplitude levels, these are the preictal, postictal, ictal and interictal stages as discussed in Deivasigamani's paper [5]. Preictal stage is related to before the seizure occurs, the ictal stage relates to the onset of an attack and the postictal follows this stage as discussed. Interictal stage is referred to as the first ictal occurring stage. Currently it is understood that seizures can be identified as early at the preictal stage [5]. Prediction and searching for signals that anticipate seizures takes significant periods of expert time, an expert can often take days, weeks or even months to analyse the output of EEG signals. In addition, people can often overlook events by accident as well as due to the enormity of the data itself as seen in [6] by Gajic *et al*. Due to these factors, there has been a push to incorporate automation to effectively minimise the cost and human effort involved in the process. Another point of interest to epilepsy experts, as discussed above, is the location and then spread of an event. The location and spread defines what the outcome of an event will be, with localised seizures generally manifesting in events such as absences and small localised arrests, these are often referred to as focal-point seizures. Non-localised seizures can engulf entire portions of the brain and tend to result in tonic-clonic, otherwise known as "grand mal," seizures. These result in the typical rhythmic muscular driven convulsions. Beghi discussed this in [7].

Benefits of using EEG for epileptic seizures stem from their cost, portability and show clear interpretable patterns in the frequency domain according to [8] by Shoeibi *et-al.* Therefore, this is particularly relevant as pervasive computation and the Internet of Things (IoT) continues to proliferate society, offering new automation opportunities as well as those that can potentially benefit, not only epileptics, but others that fall prey to similar neurologically driven events. This paper is structured as such, with Sect. 2 discussing related works in the field, Sect. 3 discusses the methodology of the investigation, Sect. 4 discusses the simulation results and finally conclusion and future works is discussed in Sect. 5, Overall, this report shows and discusses the results from applying the Random Forest algorithm with spectral features extracted by Fast Fourier Transform (FFT) to a dataset of EEG signals derived from epileptic seizures through electrodes placed on the scalp.

The process of taking an EEG involves the placement of conducting electrodes fixed with adhesive to the subject's scalp the number of which varies depending on how much area detail is necessary. Therefore, data can vary from set to set depending on the experiment carried out. Table 1 shows a selection of literature related to the study of random forests regarding this area, the Bonn dataset to be used and the implementation of Fast Fourier Transforms. The benefits of these implementations are that for a clinical situation involving considerable time to diagnose, this dataset uses 23 s of data in its unprocessed form from which classifications can be made, accurate results can be obtained often as high as the late nineties or even 100% as can be seen in [11] and [13]. However, an

epileptic going about their daily activity does not benefit themselves from this innovation. 23 s is a considerable period of elapsed time during a seizure event, given that time management is crucial, the air passage could be restricted due to the tongue being swallowed or bitten, and attention is deemed urgent so measures can be taken to reduce risk. The UCI/Bonn modified dataset considers a single second of time to make prompt decisions which is more relevant to the individual. Therefore, time to classify will be considered in this project in addition to accuracy, specificity etc.

Table 1. A selection of pertinent literature regarding this field

Classifier	Features	Performance (%)	Metric	Dataset
Extreme Learning Machine (ELM) and Multilayer Perceptrons (MLP) [9]	Sample Entropy	95.6	Accuracy	Bonn
Random Forest [10]	Short-time Fourier Transform (STFT), mean, energy and standard deviation	96.7	Accuracy	Bonn
Least Squares Support Vector Machine (LS-SVM) [11]	Fast Fourier Transform (FFT) and Discrete Wavelet Transform (DWT)	~100	Accuracy	Bonn
Random Forest [12]	Intrinsic Mode Functions (IMF)	98.4, 98.6, 96.4	Sensitivity, Specificity, Accuracy	Bonn
Random Forest [13]	Feature learning using L1-Penalised Robust Regression	100	Accuracy	Bonn

2 Methodology

In this paper, a detection mechanism is proposed that tends to focus on the diagnosis of seizure activity and discards classifying whether a seizure is tonic-clonic etc. The reasoning behind this is because an epileptic event, regardless of magnitude, tends to be significant event to the individual and the professional team in the greater picture. Given that a tonic-clonic seizure is a specifically profound event at any time during an epileptic's life with the condition, the other types of seizure are indeed just as relevant in long term analysis. The data regarding absences or otherwise can often indicate a change in status or a progression in the condition, if an epileptic has been on medication for a significant period, successfully managing the condition in doing so, then has even a light episode it could be cause for concern driving a shift in medication or otherwise.

These seizures can also be a pre-emptive sign of a larger seizure incoming thus giving an opportunity for proactive care rather than reactive, this could be the difference between someone surviving and succumbing to the infliction. This simplifies the classification process in of itself in terms of functionality, rendering it from a multiclass to a binary classification problem. The process of experimentation is as seen in Fig. 2.

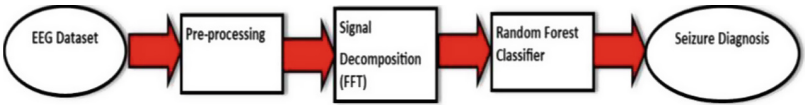


Fig. 2. Process diagram of the algorithm utilised

2.1 Dataset

Datasets are key in developing an adept learning system. An AI agent can only act as far as it has been trained properly to perceive. The demands of the dataset are dependent on several factors, application being one of them as the complexity of the system is directly related to this aspect, sitting down as still as possible in a clinical setting produces significantly different demands from a subject in motion, carrying out day to day activities. This simple caveat adds significant volumes of noise to a reading that can affect the machine learning process. Therefore, an application to be developed for this purpose of everyday diagnosis in life would need to have the relevant data, unfortunately this data is not readily available online with most readily available datasets focusing on a clinical environment application. A readily available dataset appropriate for basic seizure differentiation is the Bonn/UCI dataset as can be seen here [14]. Although it does not offer the movement data required to make a system suitable for everyday care it offers several advantages in that it has already seen data preparation to a degree for such training operations, this renders it time efficient as the pre-processing of data can take considerable time in of itself. Another aspect of simplicity is that it has been pre-rendered down to two-dimensional data, each line is a time-series datapoint which renders it simpler to interpret. The Bonn/UCI dataset also has five classes to differentiate from, these are as follows in Table 2, whilst Fig. 3 illustrates the variation between signals of each class:

Table 2. The five classifications in the Bonn Dataset [14]

Class value	Description of activity
1	Recording of seizure activity
2	They record the EEG from the area where the tumour was located
3	Yes, they identify where the region of the tumour was in the brain and recording the EEG activity from the healthy brain area
4	Eyes closed, means when they were recording the EEG signal the patient had their eyes closed
5	Eyes open, means when they were recording the EEG signal of the brain the patient had their eyes open

Given that the latter groups 2–5 correspond with non-seizure data, converting them all into class 2 is enough to turn the problem into a binary issue. This has the effect of creating a large imbalance in the dataset, a ratio of 4:1 in favour of events unrelated to seizures, this was addressed by utilising oversampling of the minority class so that the split became a 1:1 ratio. Oversampling was selected as to avoid the loss of possibly useful information provided in the majority data. Other caveats in terms of practicality in data is that epilepsy in of itself, is not merely driven by a given pattern throughout all epileptics, it is a diverse family of conditions, this means for an accurate system to be developed a large data set must be employed with as much seizure data as possible which represents are generalised tool which will not be applicable for all epileptics. Another option would be to train a highly specialised algorithm on as much individual data possible for a subject which would be a considerable project in labour and cost but would be bespoke to a given subject.

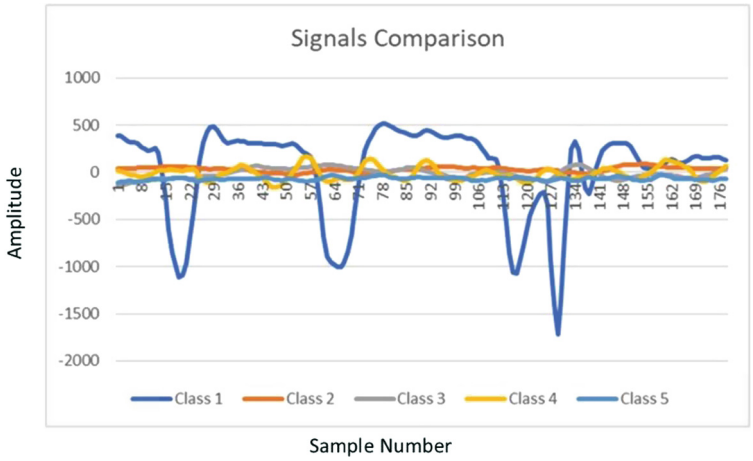


Fig. 3. Illustration of signal from each class and the differences between each

2.2 Fast Fourier Transform (FFT)

Having pre-processed the data, Fast Fourier Transform (FFT), an algorithmic technique commonly utilised in various digital signal processing applications, is applied for signal decomposition. The core principle of an FFT is to implement a fast computation of Discrete Fourier Transform (DFT) coefficients, expressing a function as a sum of periodic components using the function stated in Eq. 1 as brought forward by Cooley and Tukey in [15].

$$y[k] = \sum_{n=0}^{N-1} e^{-2\pi j \frac{kn}{N}} x[n] \tag{1}$$

where Y[k] is the FFT coefficients, N is the total number of input EEG samples, n is the signal sample index. Using this method, spectral features can be extracted from

the time series data that allows for an alternate perspective to be taken on the EEG data, effectively allowing for the time series signal to be converted into a series of components in the frequency domain. Transforms, in general, are widely implemented in the machine learning community Fourier and Wavelet transforms being readily available in addition within common libraries. Rationale for specifically choosing FFT is due to the fact it is a faster derivative of the DFT. The Discrete Wavelet Transform (DWT), an alternative algorithm utilised in this application, was not implemented because the lattice filter results in a loss of spectral resolution and, thus, can reduce the accuracy of the classification process, this can be seen in [16]. Also, given that FFT's time complexity is given by O(N*log2(N)), and DWT is given by O(N) this represents a compromise between accuracy and time, this could be justified by caveating that missing a severe seizure is riskier than the benefit from gain in processing time. The importance of a quick response within the epileptic community is well known as time can be relative to the harm caused because of seizure. The longer someone is arrested in seizure the higher the risk comes and therefore, the care needs change from something that can be handled privately to requiring an ambulance for emergency assistance. Therefore, given a second's worth of time has already been consumed to gather the necessary data for this algorithm, the actual classification process should be prompt to enable rapid attention, within seconds an individual could have collapsed and incurred concussive impact, as has been established, can result in a fatal outcome.

2.3 Random Forest Algorithm

The random forest algorithm is a form of ensemble learning that utilises several individual learners, in this case decision trees. As Safavian discusses in [17] decision tree classifiers (DTC) are pervasive in machine learning systems used in a broad range of scenarios from radar signal classification to speech recognition. The essence of DTCs is in the algorithm's ability to classify complex signals and data into simple rules, this represents a solution that is often easier to interpret. Decision Trees in of themselves can find success in time series signal data such as seizure detection with EEG signals. Random forests can offer advantages over individual learners.

As Biau *et al.* discussed in paper [18], a random forest is a general-purpose regression and classification algorithm that makes use of several decision trees arranged into an "ensemble." Figure 4 shows a simple illustration of a random forest process. This technique involves randomising these individual trees and then aggregates their individual predictions into an average score. This is a useful technique especially in areas where the variables are great in proportion to observations. Random forest has been already utilised successfully in studies regarding the analysis of seizure data. Therefore, there is already an element of verifiable proof available that this method is relatively successful within this context. In this study, 100 individual classifiers were utilised for this analysis using the standard framework given within the Scikit learn library with none of the other parameters tuned. Analytics were then carried out using a confusion matrix to understand overall performance.

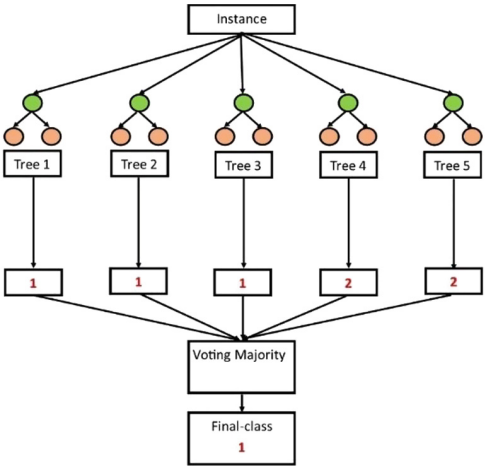


Fig. 4. A general Random Forest process

3 Results

This proposed methodology was carried out in code using Python 3.6. Within Python several libraries were utilised including SciPy for its transform suite and Scikit Learn for simple practical access to common machine learning algorithms. In addition, some pre-processing was achieved using pandas and numpy. In this experiment, the splitting and shuffling was carried out using Scikit Learn's K-fold cross-validation with shuffling engaged, and 10-fold splits being carried out amongst the data. The benefits of K-fold cross validation being a reduction in bias during learning, rendering the learner a superior generaliser. Performance was ascertained in utilisation of typical metrics in recall, precision and accuracy. As discussed, the key metric is sensitivity or recall for an epileptic seizure classifier. This operation was repeated six times to analyse if the results were consistent through multiple iterations of learning. It was then repeated on a subset of the CHB-MIT [19] dataset for parity. Table 3 shows the performance of the random forest and FFT algorithm over 6 random trials of learning and testing.

Table 3. The results of training the random forest with the FFT enabled

Trial	Recall	Precision	Accuracy
1	99.78%	98.39%	99.08%
2	100%	97.97%	98.97%
3	100%	99.73%	99.36%
4	99.78%	98.09%	98.91%
5	99.57%	97.67%	98.61%
6	99.89%	97.71%	98.79%
Average	99.84%	98.03%	98.95%

Average time to classify was 12.2 ms on 1 s of EEG data using GPU on Google Collab Cloud service. These results suggest that this methodology potentially, could lead to fast classification of epilepsy signals in a fast and timely manner when implemented on devices with low channel numbers. Of particular interest was that false negatives were minimised to the point where they were in the single digits achieving almost 100% accuracy in specificity. This means that false negatives were effectively decreased whilst training, therefore, the chance hazardous events being missed is reduced. This outperformed [9, 10] and [12] in the literature using techniques such as STFT and IMF but did not outperform [11] and [13], this could be because the dataset had more information to work with, given that these operations were based on analysing a considerable period worth of information simultaneously (23 s). Naturally, the extra information given on such a set-up would allow for alternative conclusions to be developed, possibly more accurate. These scores were achieved with one second worth of data window and 12.2 ms worth of classification time, meaning that this can detect seizures fast in the context of this dataset. Given that this was done using GPU on the Google Colab Cloud Service, it is prudent to say that if this was running on an embedded wearable device with limited resources this would run slower, further experiments could reveal the extent to which this is the case. This is far more practical for an epileptic having a seizure, reducing the time delay for a carer to act in case of an emergency. For parity, the results of a single run of the CHB-MIT were 97% accuracy, 97% precision and 97% recall. This represents a performance variation; however, this is to be expected as this subset had roughly six times as many examples than the UCI/Bonn dataset with upscaling included and 23 channels worth of information per example compared to 1. This rendered the task significantly more complex; it could be accommodated with extended work on the algorithm in terms of feature extraction for this specific digital system etc.

4 Conclusion

In conclusion, using the proposed process the system was relatively successful in diagnosing epileptic seizures from the status-quo utilising random forests and FFT algorithm achieving 99% average specificity score with a time to classify of 12.2 ms on one second of single channel data. This shows that fast, accurate seizure detection is possible. In terms of future work, it would be compelling to get more diverse data to understand if it is possible to separate epilepsy signals from movement and refine algorithms for every use by epileptic individuals. Another avenue would be to attempt proactive performance rather than reactive, to identify the seizure from the initial aura rather than when the seizure has begun. Attempting to minimize the time needed to make an accurate detection would be another experiment worth attempting.

References

1. Thijs, R., et al.: Epilepsy in adults. Lancet **393**(10172), 689–701 (2019)
2. England, M., et al.: Epilepsy across the spectrum: promoting health and understanding: a summary of the Institute of Medicine report. Epilepsy Behav. **25**(2), 266–276 (2012)

3. Fisher, R., Cross, J., French, J., et al.: Operational classification of seizure types by the International League Against Epilepsy: position paper of the ILAE Commission for classification and terminology. Epilepsia **58**, 522–530 (2017)
4. Empatica (2022, April 13). https://www.empatica.com/en-gb/index.html
5. Deivasigamani, S., Senthilpari, C., Yong, W.: Machine learning method-based detection and diagnosis for epilepsy in EEG signal. J. Ambient. Intell. Humaniz. Comput. **12**(3), 4215–4221 (2021)
6. Gajic, D., et al.: Classification of EEG signals for detection of epileptic seizures based on wavelets and statistical pattern recognition. Biomed. Eng. Appl. Basis Commun. **26**(2), 1450021 (2014)
7. Beghi, E.: The epidemiology of epilepsy. Neuroepidemiology **54**(2), 185–191 (2020)
8. Shoeibi, A., et al.: Epileptic seizures detection using deep learning techniques: a review. Int. J. Environ. Res. Publ. Health **18**(11), 5780 (2021)
9. Logesparan, L., Rodriguez-Villegas, E., Casson, A.: The impact of signal normalization on seizure detection using line length features. Med. Biol. Eng. Comput. **53**(10), 929–942 (2015)
10. Wang, X., et al.: Detection analysis of epileptic EEG using a novel random forest model combined with grid search optimization. Front. Hum. Neurosci. **13**, 52 (2019)
11. Al Ghayab, H.R., Li, Y., Siuly, S., Abdulla, S.: Epileptic seizures detection in EEGs blending frequency domain with information gain technique. Soft. Comput. **23**(1), 227–239 (2018)
12. Sharma, R., et al.: Automated system for epileptic EEG detection using iterative filtering. IEEE Sens. Lett. **2**(4), 1–4 (2018)
13. Hussein, R., et al.: Robust detection of epileptic seizures based on L1-penalized robust regression of EEG signals. Exp. Syst. Appl. **104**, 153–167 (2018)
14. UCI Machine Learning Repository. Epileptic Seizure Recognition Data Set [Online] (2017, May 24). https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition
15. Cooley, J., Tukey, J.: An algorithm for the machine calculation of complex Fourier series. Math. Comput. **19**(90), 297–301 (1965)
16. Edwards, T.: Discrete Wavelet Transforms: Theory and Implementation, Universidad De, pp. 28–35 (1991)
17. Safavian, S., Landgrebe, D.: A survey of decision tree classifier methodology. IEEE Trans. Syst. Man Cybern. **21**(3), 660–674 (1991)
18. Biau, G., Scornet, E.: A random forest guided tour. TEST **25**(2), 197–227 (2016)
19. Goldberger, A., et al.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)

An Investigation into Routing Protocols for Real-Time Sensing of Subsurface Oil Wells

Craig Stewart(✉), Nazila Fough, and Radhakrishna Prabhu

School of Engineering, Robert Gordon University, Aberdeen, UK
c.stewart35@rgu.ac.uk

Abstract. Pervasive computing has transformed society, and there is a desire to extend this mass data connectivity to the ocean, implementing an Underwater Internet of Things (UIoT), especially by energy companies seeking real-time sensor data from assets such as oil wells and pipelines. As evidenced by the Deepwater Horizon, Piper Alpha, and other disasters, failure of these assets can result in disaster. To avoid these risks, energy companies are interested in using Underwater Wireless Sensor Networks (UWSN) to achieve real-time asset monitoring, allowing for proactive maintenance. Generally, acoustic transmission technology is utilised to communicate with emerging ad-hoc UWSN, an established technology characterised by large coverage areas and reliable connectivity at the expense of high energy consumption and low operational bandwidth. Given that it is impossible to increase the speed of sound without altering the underwater channel itself physically, maximising end to end delivery time in each scenario is largely dependent on the hardware design involved and the selected protocol on the network and data link layers as well as the physical topology of the network. This simulation driven investigation aims to establish how routing technique and topology choice effects end-to-end delivery times in populated, active deep water oil drilling areas. The simulation was carried out in NS-3/Aquasim-NG and ascertained that a layered topology of fixed position nodes with Depth Based Routing (DBR) would be optimal for time critical scenarios achieving the best time between sink and source and therefore the best option for a quick response to a hazard when compared to Hop-to-Hop Vector Based Forward (HH-VBF).

Keywords: Underwater Sensor Networks (UWSN) · Routing techniques · IOT · UIOT

1 Introduction

Pervasive computation and densely connected networks are steadily encroaching into the industrial sphere. Emerging technological paradigms such as 6G Networking and Digital Twins merging with industry is a new trend in the research domain [1]. As such, this can be extended to industries with assets in the sea such as the energy or mining sectors with academia referring to new paradigms like Underwater Wireless Sensor Networks (UWSN) and the Internet of Underwater Things (IoUT) as Qiu discusses in [2]. These concepts have the potential to revolutionise activity in the subsurface domain as it

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 689–699, 2023.
https://doi.org/10.1007/978-3-031-16281-7_65

is being steadily proliferated with industries that stand to gain from exploiting formerly inaccessible resources with emerging oceanic mining on the forefront of this. Venditti of Visual Capitalist has written about ocean mining in [3]. This has a direct impact on the oil and gas industry, which has a vested interest in the potential of underwater wireless communication. This is since long-distance deep-sea communication is still carried out via tether, typically fibre optic, resulting in a cable that can be kilometers long and subject to forces at sea as well as interference from sea life such as biting sharks. There are also possible benefits such as weight shedding (the cable of an ROV is an expensive and heavy piece of equipment in of itself) and cost cutting especially as operating depth increases raising pressure and length demands along with it. Acoustic communication is still the standard for wirelessly communicating underwater due to the nature of long propagation range and relative reliability of wireless communication, propagating omnidirectionally and through solid objects, as discussed in [4] by Stojanovic and Preisig. However, because of the limited operational bandwidth, this communication methodology is incapable of transmitting multimedia such as video.

However, despite this, acoustic technology could still be of use as a signaling method within the UIoT enabling a level of autonomy. The energy industry, as well as any other industry, stands to benefit greatly from the autonomy gained through pervasive computing and the nature of proactive asset monitoring as Motlagh discusses in [6]. Proactive monitoring of assets allows for issues to be detected in the current sense rather than allowing them to develop into a risk in the future. As can be seen from a brief history of oil rig disasters, this is still a risky profession with explosions occurring every few years on a global basis resulting often in multiple deaths as well as both economic and ecological damage beyond that. Often, these tragic events are a result of an accumulation of errors, however, there are common themes that emerge, warnings were ignored, maintenance was neglected, the report [5] was a Public Inquiry into Piper Alpha by Lord Cullen, discussing the events and failing that caused the tragedy. [7] by Aasalem *et al.* discusses various academic investigations into utilising Wireless Sensor Networks (WSN) in upstream oil and gas production. This information could be transmitted through either an ad-hoc network for inspection purposes or as part of a permanent network for long term monitoring and alerting. Other purposes could be to facilitate an "Automatic Disconnection of Supply" control system where, in the case of a transient subsurface event, a sensor can propagate a command signal that automatically stops oil and gas production quickly at the emergency valves or the wellhead in the blowout prevention system. There were several issues of failure in BP's own Deepwater Horizon Report [8] that could have been identified with proactive monitoring of the safety assets facilitated by IoUT (the flat battery on the blowout preventer) or avoided completely with effective automation (data science driven risk interpretations and diagnostics on components such as the blowout preventer). Table 1 shows sensor technologies utilised for the inspection subsurface pipelines.

This investigation aims to carry out research into the current field of UWSN technologies to investigate practical network paradigms that will allow emergency systems to be proactively maintained via diagnostics so that issues can be identified and solved before there can be accumulation of risk factors resulting in a catastrophic event such as Deepwater Horizon. If that fails, then this same network can be used to transmit

low data command signals that can operate emergency systems. Having identified the relevant technology, a simulation in NS-3/Aquasim-NG [9, 10] respectively will be carried out that shows that the novel UWSN network can carry out this function. This investigation will cover the following subjects organised in the following sections in this report. Section 2 covers the relevant work into the real-time monitoring of Subsea assets. Section 3 describes the theoretical network parameters and Sect. 4 describes the carried-out simulation as based on these parameters. The results and discussion of this investigation will be brought forth in Sect. 5 with the conclusion and future work discussed in Sect. 6.

Table 1. Typical sensors used for inspecting pipelines [11]

Issue	Technologies
Erosion	Acoustic Emission
	Electromagnetic Acoustic Transducer
	Computed Tomography
	Remote Field Eddy Current
	Ultrasonic sensor technologies
Corrosion	Electromagnetic Acoustic Transducer
	Remote Field Eddy Current
	Computed Tomography
	Guided Ultrasonics
	Magnetic Flux Leakage
Fatigue	Electromagnetic Acoustic Transducer
	Sonar Imaging
	Fibre Optic Deformation sensors
	Remote Field Eddy Current
	Computed Tomography
Deformation	Fibre Optic Deformation
	Sonar Imaging
	Motion Detection technologies

2 Real-Time Monitoring of Subsea Assets

There are not many case studies within the academic body for real time monitoring the subsea oil and gas assets, most technological work tends to focus on generic situation with non-descript functions being carried out. A case was identified where FJL Ribeiro *et al.* discussed the Campos Basin in Brazil [12]. This lack of publications suggests that there is a gap that needs to address regarding how these technologies could be implemented

in industrial scenarios. Given this, the literature review identified several case studies to investigate for root causes that can be addressed with UWSN technology. Although, there have been many oil production disasters, two of the most infamous were the Piper Alpha (1988) and Deepwater Horizon (2010) incidents which both resulted in mass casualties and landed the operators with billions in fines. Reports detailing the findings of the incidents are given by [5] and [8] respectively. In both reports there were incidents propagating that could be used in modern systems to control the process automatically, reducing the impact of the disasters. From there, feasible geographic locations were picked based on the number of oil rigs present in that area for the experiment to be based on. These were the North Sea (184), the Gulf of Mexico (175), the Persian Gulf (159) and Far East Asia (155) according to the data [13] from Statista. Physical layer technology was then identified that could facilitate transmission of low-data rate sensor data through the aquatic channel, the series of 42/65 modem from EvoLogics was selected to define the parameters used in the simulation. This series comes as a diverse range of modern devices that allow for an Underwater Acoustic Sensor Network (UASN) to be realised. They can be sourced from the EvoLogics website [14]. The next task was to research the routing algorithms readily available to be utilised in the simulation. The protocols chosen were Depth Based Routing (DBR) [15] and Hop to Hop Vector Based Forwarding (HH-VBF) [16]. These protocols are famous and readily available, a lot of modern routing technologies are based on these founding protocols. These protocols have also seen regular development since these initial papers [17].

3 Case Study

The key issue of this simulation is to produce a simulation that shows a network that can quickly deliver diagnostic data from sensor or send control data to valves from the seabed when a risk is detected. Depending on the depth of the water and type of media being transmitted this can have ramifications on network design as in an acoustic network as distance and bandwidth parameters are in direct conflict in relation to one another, a caveat must take place to achieve significant capacity for sensor data whilst achieving the range to reach the surface. These ideas are discussed in [4]. Depths were picked as to reflect the deepest wellhead in each field (to communicate wirelessly from inside the well itself would require a different physical layer technology), as such, it can be seen in Table 2 that there is a large variation in depths.

Table 2. Depths of some select Deepwater energy projects

Location of well	Depth of wellhead
Rockall Trough (North Sea)	1886 m
Perdido (Gulf of Mexico)	2450 m
Salman (Persian Gulf)	43 m
DeepSea 1 (Far East Asia)	1500 m

Given that 42/65 series modems can cover 1000 m transmission distance with a maximum 31.2 kbits/s data rate the deeper wellheads will require relay nodes between the sea surface and the seabed to extend the range. This bit rate will be ample enough to allow for seismic sensors to propagate data as these devices are generally low sample rate due seismic waves being predominantly low frequency in nature. This results in different topologies being implemented for each wellhead. Figure 1 shows an example of a novel Deepwater acoustic sensor network [12]. This fig shows how sensors and devices in the seabed can communicate data via acoustics to surface vessels, forming a component of the IoUT. These vessels can then communicate the aggregated data via typical Wide Area Network (WAN) technology such as radio frequency (RF) communications. This allows for the data to be integrated with the greater Internet of Things. Data and alerts could be feasibly transmitted out to experts and emergency services onshore within a few seconds of data being propagated at the wellhead. This would allow for quicker responses than what has typically been seen in disaster scenarios. Given this the four networks were to be designed in such a way that the device at the wellhead, whether it is a blowout preventer or a seismic sensor, can communicate with the surface, there must also be several paths that can be taken to reduce the impact of a single relay node failing. Therefore, the networks took the following 3D grid-based shape based on this information.

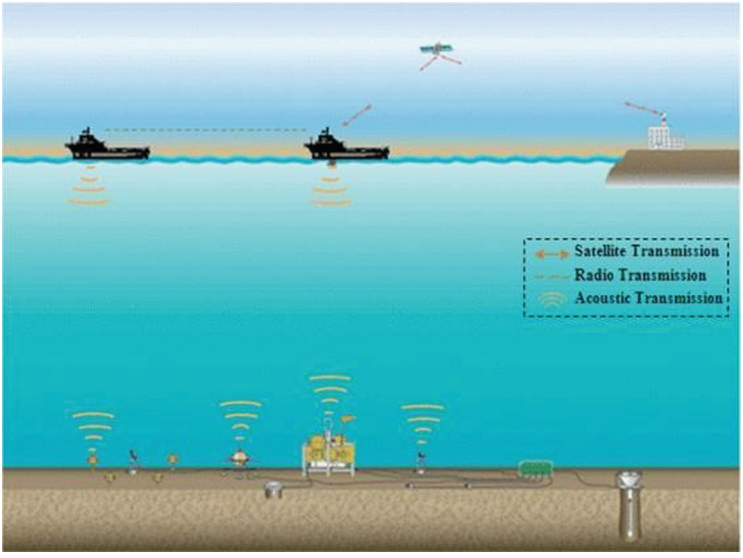


Fig. 1. A visual of a basic UASN [12]

Figure 2 shows a visual of how the nodes would be arranged to transfer the data from seafloor to sea-surface dependent on the depth of the wellhead in relation to the surface. As can be seen, the network varies in complexity depending on the depth of the wellhead at the location. The oil field at Salman can function on a relatively simple network as the Persian Gulf is relatively shallow compared to many deep-sea operations reaching only

tens of meters at its deepest point. The other three networks are relatively complex with varying numbers of nodes and layers being utilised. These layers are spaced out so that they are within the 1000 m range of the source and sink node's acoustic modem, allowing for reliable data transfer to take place from source to sink. The purpose in having the layered approach with multiple nodes per layer rather than just a single node is to allow for the redundancy in the routing to take place. These layered nodes are anchored in a square formation whose sides are 100 m length, with a single node in the middle. The reason for this, is to offer distance between the asset and the critical communication nodes, so that if debris falls and knocks out nodes automated emergency systems can still activate, limiting damage. In this situation, if all nodes were clustered in proximity and debris collapsed down on that location there is a significant chance that the flow of communication could be disrupted, preventing emergency system from activating. The routing technique was selected as DBR for numerous reasons regarding the case study, as discussed, it is computationally simple and processes the data quickly, greedily forwarding the packeted data onwards with minimal calculations taking place. It is also useful as it is decentralised from a central hub using only pressure information to make routing decisions, this renders the network like a mesh topology, a robust formation where there is no single point of failure in routing. Therefore, in an emergency, if the network is maintained in terms of battery and diagnostics are carried out frequently, whether automatically or manually, this network will be robust enough to transmit sensor or command data when necessary and, of prime concern, in an emergency scenario. HH-VBF was selected for similar reasons, it is a simple computationally, decentralized algorithm with a high delivery ratio in sparse UWSN, it is also a standard protocol in the UWSN literature when it comes to drawing comparisons overall. Once again rendering the network into a mesh topology with no single point of failure. This protocol forwards data based on a calculated vector between the source and the sink, then establishes routes on a hop-to-hop basis based on this vector. The network should be, therefore, relatively robust.

In this scenario, it would be desirable to break the chain reaction of accumulating problems as quick as possible considering the gravity of these disasters. Although the events themselves take minutes or hours from onset to eventual complete collapse, the time in between an individual incident leading to a fatal problem itself can be quick, as seen in the reports [5] and [8]. Unfortunately, acoustic signaling is relatively slow in comparison to RF and optical propagation (acoustics propagate at 1500 m/s in water whereas optics and RF propagate at $3 * 10^8$ m/s). This means that there will always be a finite delay in time between actuality and perception particularly at longer distances. What can be affected however, is the processing time, this is influenced by the number of nodes on a given route and routing process itself. Therefore, a well-designed network can limit this time to be as small as possible. The underwater acoustic networks, as seen in Fig. 2, have been designed to keep this in mind. Concerns about power can be erased or alleviated depending on how the network is implemented and how the applications are operated. If the application is managed as such that it is constantly being accessed, then battery will be consumed excessively due to the nature of acoustics requiring such power to transmit. However, if the network design is to be utilised sporadically with a wake-up scheme that means it activates only when needed then the network can be preserved for

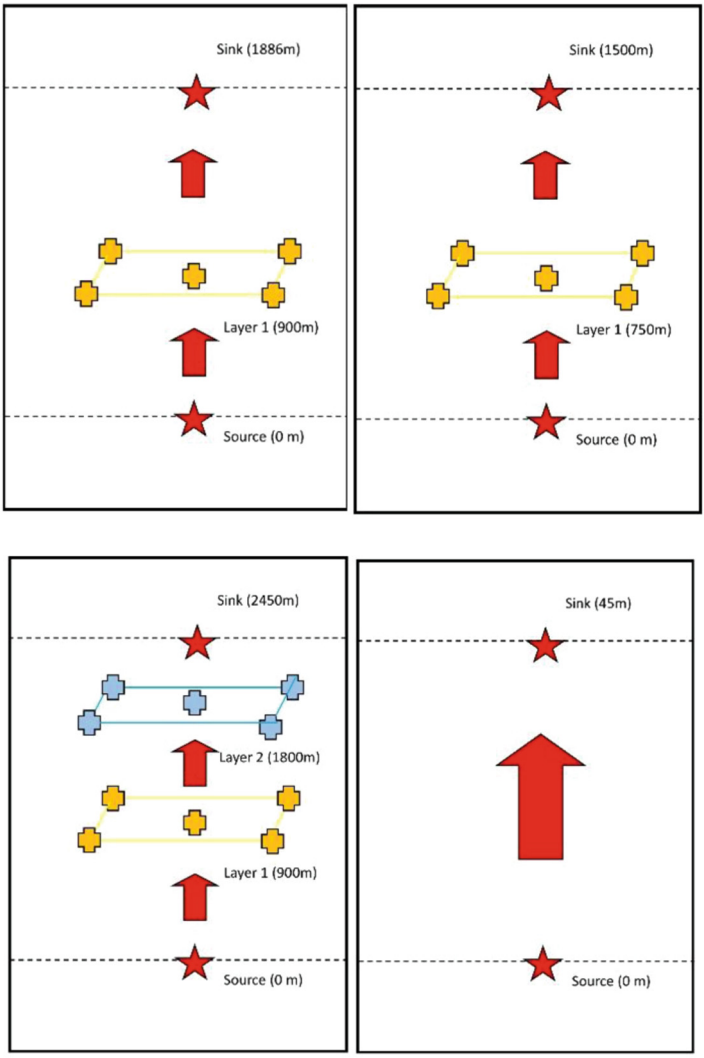


Fig. 2. Network topology according to the different location and depths. From left to right, Rockall Trough, Deep Sea 1, Perdido and Salman

a considerable amount of time. Thus, if this is to be the case, for this simulation at least, lifetime is not considered.

4 Simulation Results

A simulation was carried out utilising Network Simulator 3 (NS-3) and the UASN extension library Aquasim-NG. NS-3 is an open-source C++ library for simulating all manners of experimental networks and it is commonly utilised in research throughout the academic community. This software was utilised to implement the scenarios above and test for connectivity then end to end delay. Tracing packets will confirm that the system can establish end-to-end transmission and the length of time needed to reach the surface. The parameters for the simulation were as follows in Table 3.

Table 3. The parameters used in the simulation

Parameter	Value
Source Node	1
Sink Node	1
Relay Nodes	Variable
Depth	Variable
Communication Range	1000 m
Data Rate	31.2 kb/s
Packet Size	400 bits
Speed of Sound	1500 m/s
HH-VBF Pipe Width	100 m

Results can be analysed by confirming from the generated the trace file that a packet generated by the source has reached the designated sink through the series of relay nodes, then using Formula 1 the end-to-end transmission time can be determined.

$$End\ to\ End\ Time = Trx - Ttx \tag{1}$$

T_{rx} symbolising the time instant in seconds where the sink receives the first copy of a source generated packet and T_{tx} being the time instant where the packet is first propagated in the network from the source.

Formula 2 describes how propagation delay can by calculated. This metric being the network induced time delay. This is given as: -

$$Propagation\ Delay = End\ to\ End\ Time - \frac{Distance\ between\ Source\ and\ Sink}{Speed\ of\ Sound} \tag{2}$$

The results are as follows as displayed in Table 4 and Fig. 3.

Table 4. Results of each network simulation

Location	End to End Connectivity	DBR End to End Time (Secs)	HH-VBF End to End Time (Secs)	DBR Propagation Delay (Secs)	HH-VBF Propagation Delay (Secs)
Rockall Trough	Yes	1.368	2.844	0.114	1.590
Perdido	Yes	1.817	3.756	0.184	2.123
Salman	Yes	0.091	0.825	0.062	0.796
DeepSea	Yes	1.111	2.047	0.111	1.074

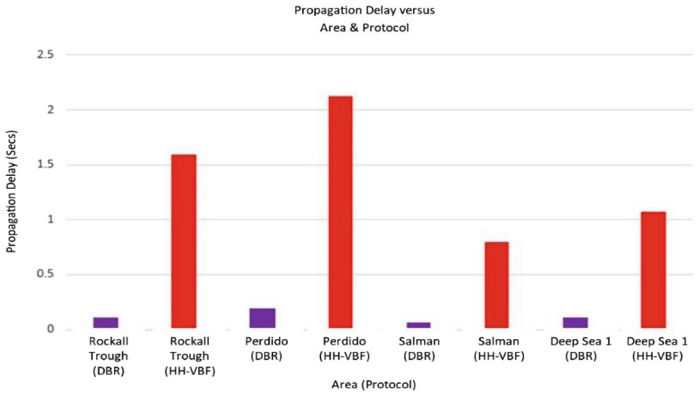


Fig. 3. A comparison of propagation delay versus protocol in four different areas

As can be seen in Table 4, DBR was proven to operate the quickest of the available routing techniques, in this scenario. DBR was quick to process packets consistently reaching the surface within less than 0.2 s processing time within the network, reaching the destination. HH-VBF, however, took considerably more time in those fronts with propagation delays of beyond a second within the same networks. This means that DBR would be preferable for this formation of network as it is consistently quicker than HH-VBF. The speed overall for DBR would be great for these types of services. As can be seen in the reports above, there is a mix of times between problems emerging then manifesting physically. As such, given this robust, quick network response, it is likely that if this network was to have been implemented in one of the given scenarios it would have allowed for interjection quickly at given stages throughout the chain of events and allowed for ample opportunities to prevent disaster. This can be said because the events that unfolded generally took a finite amount of time to propagate between several seconds and minutes, a quick response at one or more of those points could have altered the course of these disasters. Naturally, the quicker the network response and lower the propagation delay, the more chances open from which events can be interrupted.

5 Conclusion

In conclusion, this study effectively demonstrated through simulation how a network could be deployed that allows for sensing and quick or automatic disconnection at the wellhead. A literature review was carried out that looked at disasters such as Piper Alpha and Deepwater Horizon, identifying stages where the deployment of UWSN could have helped carry out proactive maintenance or directly allowed for interjection into the chain reaction that led to the disasters. A 3D layer-based approach to positioning acoustic nodes with a mesh topology was taken to the problem that allowed for nodes to distributed according to the operating range of the modems. This design allowed for the effective range of a modem positioned at the wellhead to be extended to the surface. It was found that data can be transferred quickly, in a matter of seconds, dependent on distance. It discussed how this paradigm could allow for quicker manual or automatic decision making to be made in emergencies, feasibly allowing for valves and blowout protection to be activated remotely in an emergency. This extra input could theoretically be the difference between a small problem remaining solvable or ballooning into a catastrophe like Piper Alpha. Future works considered could be looking into developing an algorithm that reliably identifies when a disaster is possible for this system.

References

1. Khan, L.U., et al.: Digital-twin-enabled 6G: vision, architectural trends, and future directions. IEEE Commun. Mag. **60**(1), 74–80 (2022)
2. Qiu, T., et al.: Underwater Internet of Things in smart ocean: system architecture and open issues. IEEE Trans. Industr. Inf. **16**(7), 4297–4307 (2019)
3. Deep Diving for Metals: Visualizing Ocean Mining, Visual Capitalist, August 2021. https://www.visualcapitalist.com/deep-diving-for-metals-visualizing-ocean-mining/
4. Stojanovic, M., Preisig, J.: Underwater acoustic communication channels: propagation models and statistical characterization. Commun. Mag. IEEE **47**, 84–89 (2009)
5. WD Cullen: Public Inquiry into the Piper Alpha Disaster. Department of Energy, London, UK, Cm 1310 (1990)
6. Motlagh, N.H., Mohammadrezaei, M., Hunt, J., Zakeri, B.: Internet of Things (IoT) and the energy sector. Energies **13**(2), 494 (2020)
7. Aalsalem, M.Y., et al.: Wireless Sensor Networks in oil and gas industry: recent advances, taxonomy, requirements, and open challenges. J. Netw. Comput. Appl. **113**, 87–97 (2018)
8. Deepwater Horizon Investigation Report, BP (2010)
9. nsnam, 21 March. https://www.nsnam.org/
10. Martin, R., Rajasekaran, S., Peng, Z.: Aqua-sim next generation: an NS-3 based underwater sensor network simulator. In: Proceedings of the International Conference on Underwater Networks & Systems, pp. 1–8 (2017)
11. Davis, P., Brockhurst, J.: Subsea pipeline infrastructure monitoring: a framework for technology review and selection. Ocean Eng. **104**, 540–548 (2015)
12. Ribeiro, F.J.L., de Castro Pinto Pedroza, A., Costa, L.H.M.K.: Underwater monitoring system for oil exploration using acoustic sensor networks. Telecommun. Syst. **58**(1), 91–106 (2014). https://doi.org/10.1007/s11235-014-9948-6
13. Sönnichsen, N.: Number of offshore rigs worldwide as of January 2018 by region. Statista, January 2018. https://www.statista.com/statistics/279100/number-of-offshore-rigs-worldwide-by-region/

14. Acoustic Modems. EvoLogics, March 2022. https://evologics.de/acoustic-modems
15. Yan, H., Shi, Z.J., Cui, J.-H.: DBR: depth-based routing for underwater sensor networks. In: Das, A., Pung, H.K., Sung, F.B., Lee, L.W., Wong, C. (eds.) NETWORKING 2008. LNCS, vol. 4982, pp. 72–86. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-79549-0_7
16. Nicolaou, N., et al.: Improving the robustness of location-based routing for underwater sensor networks. In: OCEANS 2007, Europe, pp. 1–6 (2007)
17. Luo, J., et al.: A survey of routing protocols for underwater wireless sensor networks. IEEE Commun. Surv. Tut. **23**(1), 137–160 (2021)

Surface-Functionalized Multichannel Nanosensors and Machine Learning Analysis for Improved Sensitivity and Selectivity in Gas Sensing Applications

Luis Antonio Panes-Ruiz[1], Shirong Huang[1], Leif Riemenschneider[1], Alexander Croy[2], Bergoi Ibarlucea[1,3](✉), and Gianaurelio Cuniberti[1,3](✉)

[1] Institute for Materials Science, Max Bergmann Center for Biomaterials, Dresden University of Technology, 01062 Dresden, Germany
{bergoi.ibarlucea,gianaurelio.cuniberti}@tu-dresden.de
[2] Institute for Physical Chemistry, Friedrich Schiller University, 07743 Jena, Germany
[3] Center for Advancing Electronics Dresden (cfaed), Dresden University of Technology, 01062 Dresden, Germany

Abstract. Breath analysis is an emerging technique in the field of diagnostics. The presence of thousands of gases and volatile organic compounds (VOCs), many of them at part per billion (ppb) concentration levels, require the development of ultrasensitive and selective detection approaches, which pose challenges still trying to be addressed by the scientific community. Here, we describe two approaches that provide a substantial contribution to the development of gas sensors. The first one is based on modifications of the used sensing material, namely a specific surface functionalization based on gold nanoparticles of carbon nanotubes to achieve selectivity toward hydrogen sulfide, together with the implementation of multiple sensors for self-validation. The second one focuses on the analysis method, implementing machine learning algorithms to maximize the information obtained from each single sensor to distinguish gases based on their interaction kinetics with the sensor. The combination of both approaches is foreseen as a powerful tool for the development of new smart sensing platforms with high potential in terms of analytical efficiency.

Keywords: Gas nanosensors · Breath diagnostics · Machine learning

1 Introduction

Exhaled breath analysis has emerged during the last decade as a promising tool for non-invasive and cost-effective disease identification and health monitoring [1, 2]. The complexity of breath imposes a serious challenge considering the thousands of gases and volatile organic compounds that it contains [3], many of them at extremely low levels in the order of ppb. Hence, among the existing bottlenecks to obtain a diagnosis with high precision, selectivity and sensitivity are prominent ones. Together with them

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 700–707, 2023.
https://doi.org/10.1007/978-3-031-16281-7_66

we may mention stability, reproducibility, fast response time, and miniaturization possibility for portable analysis consuming low energy. In this context, the implementation of nanomaterials in the low-power gas sensor fabrication using micro- and nanofabrication processes addresses the miniaturization and sensitivity issues [4]. Selectivity is an issue with a more difficult solution. This is typically addressed following specific functionalization strategies that make the device more prone to react with the target gas [5]. However, in contrast to biosensors, where the millions of years of evolution has given as result the existence of bioreceptors of exquisite specificity [6], functionalized gas sensors still offer only partial selectivity. Complementing the used hardware with machine learning algorithms or using neural networks helps to improve the selectivity to certain extent [7, 8].

Here we describe two approaches that address both sensitivity and selectivity issues and offer a potential to be combined in a cooperative way toward improved gas sensor performance. First, we modify semiconducting carbon nanotubes-based chemiresistive sensors with gold nanoparticles and implement them in multichannel arrays for self-validation [9]. The modification provides the sensor with an improved selectivity and sensitivity toward H_2S as marker of small intestinal bacterial overgrowth [10] due to the affinity between gold and sulfur. Second, we introduce the use of machine learning algorithms for the identification of gases with single-channel sensors [11], based on the different adsorption and desorption kinetics, analyzing the transient features during exposure and flushing phases.

2 Selective Detection of H_2S Gas by AuNP Functionalized Carbon Nanotubes

2.1 Multichannel Gas Sensing Device Fabrication

Multichannel devices of 2.4 × 2.4 cm were fabricated on Si/SiO_2 by standard lithography technique and deposition of Cr and Au. The device consisted of 64 individual interdigitated electrode (IDE) areas of 5 μm in width and gap size of 2 μm arranged in 4 groups of 16, sharing a common source/working electrode used for nanotube deposition and gold nanoparticle (AuNP) functionalization (Figs. 1a and 1b). The semiconducting single-walled carbon nanotubes (sc-SWCNTs) were specifically deposited onto the IDE areas by a controlled dielectrophoretic deposition process. Briefly, an N-methyl-2-pyrrolidone (NMP) based sc-SWCNT dispersion was placed on the device and an AC electric field of 8 V_{pp} and 5 kHz was applied sequentially to each channel (IDE structure) while monitoring the electrical resistance change. The process was stopped when the desired resistance value was reached, thus ensuring a narrow distribution of the sensor's final electrical resistance in the low kΩ range. This technique allowed the alignment of individual and small sc-SWCNT agglomerations between the electrodes which provided an increased active surface area for the interaction with the target gas molecules (Figs. 1c and 1e).

To enhance the selectivity toward H_2S gas, we took advantage of the high chemical affinity of gold and sulfur by functionalizing the sc-SWCNTs with gold nanoparticles (AuNP) using a potentiostatic electrodeposition process and 1mM $KAuCl_4$ as electrolyte.

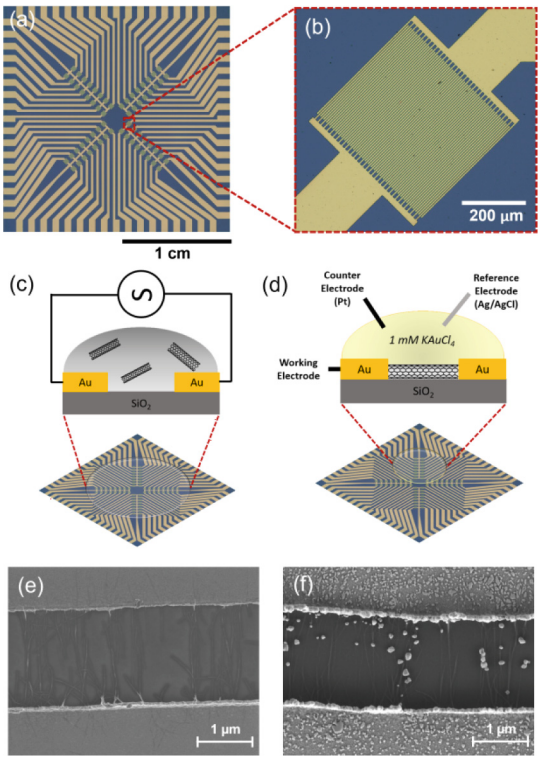


Fig. 1. Multichannel gas sensing device fabrication. (a) Photograph of the multichannel device with 64 gas sensors diagonally distributed in 4 groups of 16. (b) Photograph of an individual interdigitated electrode (IDE) area with gaps of 2 µm. (c) Schematic representation of the dielectrophoretic deposition process of sc-SWCNTs. (d) Schematic representation of the AuNP electrodeposition process. (e) Scanning electron micrographs of aligned sc-SWCNTs between metallic electrodes after dielectrophoretic deposition and (f) after AuNP functionalization. Modified with permission from [9] under a Creative Commons Attribution 4.0 International License.

Briefly, the electrolyte was placed on the device covering one of the four groups at a time. Then, the counter (platinum wire) and reference (Ag/AgCl wire) electrodes were immersed and a constant potential of −0.4 V was applied for 30 s. This process led to the nucleation of AuNP at the nanotube's defect sites and nanotube-nanotube junctions which are more chemically reactive [12, 13] achieving an average particle diameter of 60 nm and separation distances of at least 100 nm (Figs. 1d and 1f).

2.2 Gas Sensing Performance

To investigate the effect of AuNP functionalization on the sensing response to H_2S and NH_3 gases, AuNP-functionalized and non-functionalized sensors were exposed to 4 cycles of 20, 40, 80 and 160 ppb of H_2S and NH_3 at room temperature and relative humidity of 25% using air as carrier gas. The sensors were tested in a chemiresisitive format by applying a constant source-drain voltage (V_{SD}) of 100 mV and measuring the

change in electrical current upon exposure to the target gases. The sensing response was then calculated as the relative resistance change using the following equation [14]:

$$Response(S)(\%) = \frac{\Delta R_{(t)}}{R_0}$$

As can be seen in Fig. 2a and b, AuNP-functionalized sensors showed a higher sensing response to all H_2S concentrations compared to non-functionalized ones. Also, a similar response can be observed from all sensors over the four continuous cycles demonstrating high reproducibility and self-validating capabilities. Moreover, the response shows good linearity ($R^2 = 0.98$), a sensitivity of 0.122%/ppb and a calculated limit of detection of 3 ppb. The enhanced sensitivity comes from the interaction of H_2S molecules at the AuNP/nanotube interface producing a catalytic spillover effect injecting electrons from the AuNP to the sc-SWCNTs, thus increasing the electrical resistance of the system (Fig. 2a and 2c) [15, 16]. On the other hand, a similar response from both types of sensors was observed in all NH_3 gas concentrations. Here, the effect at the AuNP-nanotube interface does not take place and the main contribution to the resistance change is the Schottky barrier modulation at the nanotube-electrode interface present in both types of sensors (Figs. 2b and 2d).

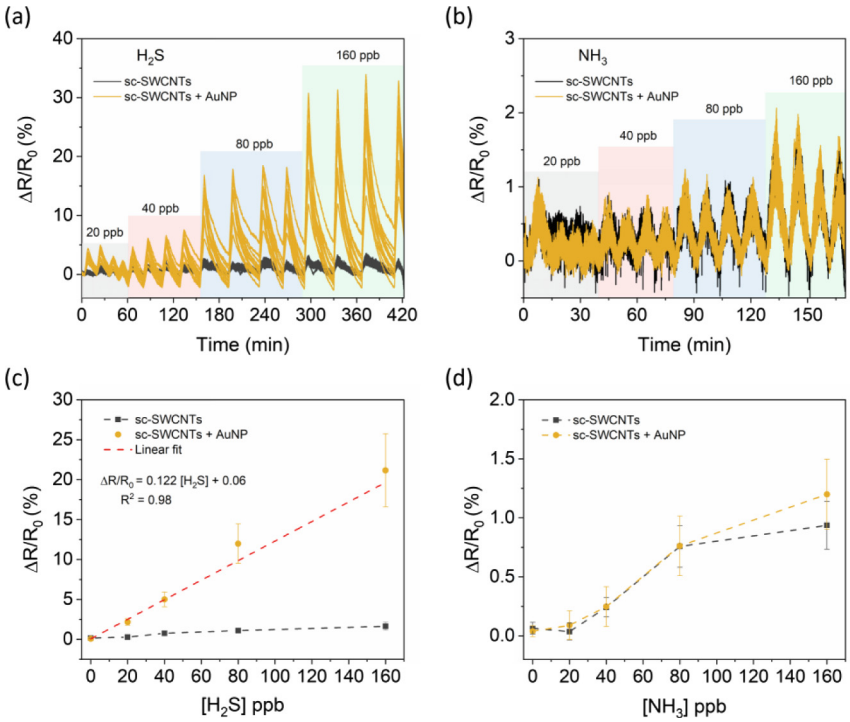


Fig. 2. Sensing response of non-functionalized and AuNP functionalized sensors to ppb concentrations of H_2S and NH_3. Adapted with permission from [9].

3 Machine Learning-Enabled Gas Identification

Next, we analyzed the signal response using machine learning algorithms to maximize the outcome of single channel sensors, aiming at identifying gases with the minimum possible hardware. Here, we selected NH_3 and PH_3 as target gases with similar properties, using N_2 as reference gas, as proof of principle.

The time dependent response signals (Fig. 3a) consisted of 24 cycles of two phases each: analyte gas exposure phase and analyte gas flushing phase. In order to obtain the gas features for classification, multiple transient features were extracted (Fig. 3b). The procedure of feature extraction was as follows: first, time-dependent current was transformed into time-dependent resistance. Then, the 24 individual measurement cycles were split to obtain multiple single gas analysis profiles consisting of signal stabilization, gas exposure and gas flushing. Third, the fractional change of the resistance was derived and the sensing response profile was obtained. Next, data was normalized using L2 norm algorithm.

Eleven transient features were extracted from the response profile, using a Python script [17]. The exposure (t_1-t_2) and flushing phases (t_2-t_3) were fitted by exponential functions, obtaining three coefficients (a_1, b_1, c_1) for the first phase and another three (a_2, b_2, c_2) for the second phase. The first and second derivatives as a function of time were calculated after fitting with a polynomial function, from which the maximum (k_{max}) and minimum (k_{min}) values of the first derivative and the minimum value (a_{min}) of the second derivative were determined. The transient response amplitude (S) in the exposure phase (t_1-t_2) was calculated. Additionally, the area under the whole response profile (t_1-t_3) was integrated. In total, each gas was represented by a feature vector consisting of 24 arrays of 11 features. The same procedure was followed for NH_3 and PH_3 gases as well as pure N_2 as reference at three different concentrations, 100 ppb, 500 ppb and 1000 ppb.

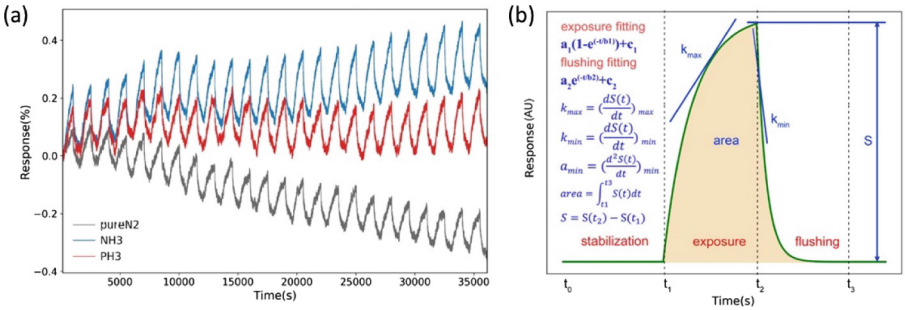


Fig. 3. (a) Gas sensing response profile towards analyte gases at 500 ppb concentration. (b) Schematic sensing response profile $S(t)$ for single cycle test, consisting of stabilization process (t_0-t_1), analyte exposure phase (t_1-t_2), and analyte flushing phase (t_2-t_3). The feature vector representing each analyte gas consists of 11 transient parameters, including, a_1, b_1, c_1, a_2, b_2, c_2, S, k_{max}, k_{min}, a_{min}, and area. Adapted with permission from [11].

The data was analyzed employing supervised machine learning (Linear discriminant analysis, LDA), obtaining the score plots shown in Fig. 4 (a–c). The LDA algorithm

attempts to find a feature subspace that optimizes class separability [18]. As it is observed, at 100 ppb concentration, NH_3 forms an isolated cluster while PH_3 cluster exhibits some overlapping with reference gas cluster. At 500 ppb and 1000 ppb concentration, the three clusters separate from each other well, which suggests a robust classification of these three analyte gases. The LDA classification results are consistent with the PCA results.

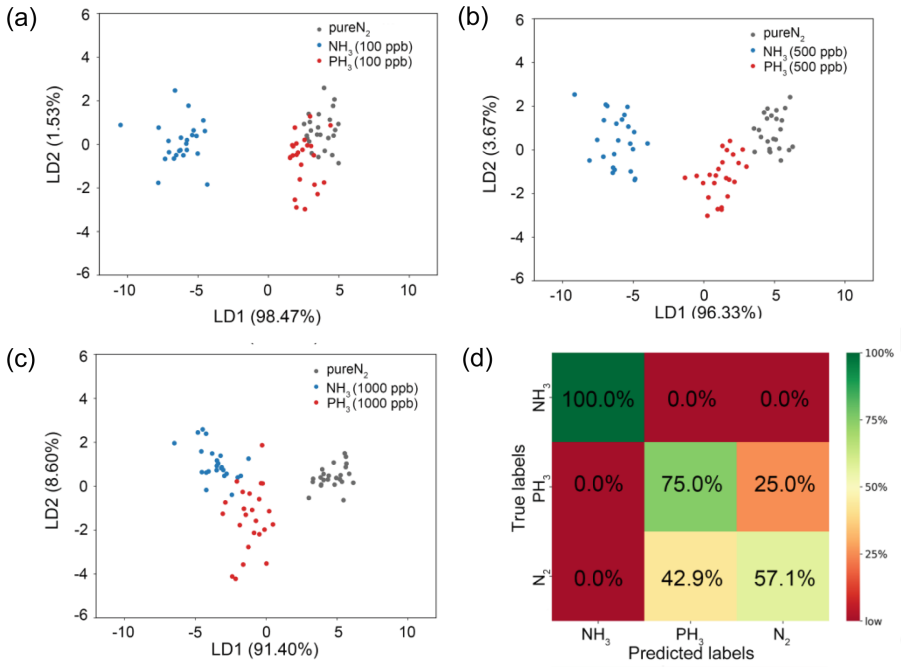


Fig. 4. LDA score plot for analyte gases at (a) 100 ppb, (b) 500 ppb, and (c) 1000 ppb concentration. (d) Confusion matrix of gas classification at 100 ppb concentration Adapted with permission from [11].

To evaluate the classification performance, a hold-out cross-validation approach was employed to calculate the confusion matrix as classification performance evaluation. The LDA classifier was trained with 70% of the feature data and the rest was applied to validate the trained classifier. Figure 4(d) shows the confusion matrix results for 100 ppb. Good classification results were achieved for NH_3, with only moderate performance for PH_3 due to the overlapping with N_2. At higher concentrations (500 ppb and 1000 ppb), the sensor presented an excellent classification performance for all gases.

4 Conclusions

We have demonstrated two approaches that separately contribute to improve the sensitivity and selectivity of gas sensors. The first one consists on tuning the hardware, from functionalization for enhanced interaction with the target gas to implementation

of multiple sensors for self-validation. The results show that the sensitivity can reach ultralow detection limits with good sensor-to-sensor reproducibility and with clear increment of the response toward the selected target gas H_2S as marker of intestinal diseases, compared to NH_3 which is also present in breath.

The second approach consists of an improved methodology for the analysis of the sensor response signal using machine learning algorithms, considering not only a single feature (response amplitude) but also several other transient features related to the adsorption and desorption kinetics of the gas on the sensor surface. Two gases with similar properties, namely NH_3 and PH_3, as well as N_2 as reference are identified by using a single sensor, evidencing the potential of the employed methodology. This approach is strongly software based and can adapt to other hardware strategies such as those based on silicon nanowires [19].

While both approaches have been employed separately, we predict that their combination as well as the integration of additional materials and functionalizations constitutes a promising tool toward a wider range of gases and more complex scenarios with gas mixtures.

Acknowledgements. The authors acknowledge the funding support from 6G-life project (Federal Ministry of Education and Reserach of Germany in the programme of "Souverän. Digital. Vernetzt.," project identification no. 16KISK001K).

References

1. Hunter, G.W., Dweik, R.A.: Applied breath analysis: an overview of the challenges and opportunities in developing and testing sensor technology for human health monitoring in aerospace and clinical applications. J. Breath Res. **2**(3), 1–7 (2008)
2. Sheridan, C.: COVID-19 spurs wave of innovative diagnostics. Nat. Biotechnol. **38**, 769–772 (2020)
3. Popov, T.A.: Human exhaled breath analysis. Ann. Allergy Asthma Immunol. **106**(6), 451–456 (2011)
4. Panes-Ruiz, L.A., et al.: Toward highly sensitive and energy efficient ammonia gas detection with modified single-walled carbon nanotubes at room temperature. ACS Sens. **3**(1), 79–86 (2018)
5. Jung, H.Y., et al.: High-performance H2S detection by redox reactions in semiconducting carbon nanotube-based devices. Analyst **138**, 7206–7211 (2013)
6. Luo, X., Davis, J.J.: Electrical biosensors and the label free detection of protein disease biomarkers. Chem. Soc. Rev. **42**(13), 5944–5962 (2013)
7. Hayasaka, T., et al.: An electronic nose using a single graphene FET and machine learning for water, methanol, and ethanol. Microsyst. Nanoeng. **6**(1), 50 (2020)
8. Al Chawa, M.M., et al.: Gas sensing discrimination using a cellular nonlinear network. In: International Workshop on Cellular Nanoscale Networks and Their Applications, pp. 1–3. IEEE Computer Society (2021)
9. Panes-Ruiz, L.A., et al.: Selective and self-validating breath-level detection of hydrogen sulfide in humid air by gold nanoparticle-functionalized nanotube arrays. Nano Res. **15**, 2512–2521 (2022)
10. Banik, G.D., et al.: Hydrogen sulphide in exhaled breath: a potential biomarker for small intestinal bacterial overgrowth in IBS. J. Breath Res. **10**(2), 026010 (2016)

11. Huang, S., et al.: Machine learning-enabled smart gas sensing platform for identification of industrial gases. Adv. Intell. Syst. **4**, 2200016 (2022)
12. Fan, Y., et al.: Identifying and counting point defects in carbon nanotubes. Nat. Mater. **4**(12), 906–911 (2005)
13. Charlier, J.C., et al.: Carbon nanotubes randomly decorated with gold clusters: from nano^{2}hybrid atomic structures to gas sensing prototypes. Nanotechnology **20**(37), 375501 (2009)
14. Battie, Y., et al.: Thickness dependent sensing mechanism in sorted semi-conducting single walled nanotube based sensors. Analyst **137**(9), 2151 (2012)
15. Mubeen, S., et al.: Sensitive detection of H_2S using gold nanoparticles decorated SWNTs. Anal Chem. **82**(1), 250–257 (2010)
16. Mubeen, S., et al.: Gas sensing mechanism of gold nanoparticles decorated single-walled carbon nanotubes. Electroanalysis **23**(11), 2687–2692 (2011)
17. Van Rossum, G., Drake, F.L., Jr.: Python 3 Reference Manual, version 3.7.3. Scotts Valley, CA (2009)
18. Sharma, A., Paliwal, K.K.: Linear discriminant analysis for the small sample size problem: an overview. Int. J. Mach. Learn. Cybern. **6**(3), 443–454 (2014)
19. Mikolajick, T., et al.: Silicon nanowires – a versatile technology platform. Phys. Status Solidi RRL **7**(10), 793–799 (2013)

Design and Simulation of a Novel Low-Voltage RF MEMS Switch for Reconfigurable Antennas

Alaa Elshazly[1], Mario Mounir[1], Muhammad K. Khalaf[1], Faycal Saffih[2], Yasmine Elogail[1], and Hassan Mostafa[1(✉)]

[1] Nanotechnology and Nanoelectronics Engineering Department, UST at Zewail City, Giza 12578, Egypt
hmostafa@uwaterloo.ca, hmostafa@zewailcity.edu.eg

[2] Optoelectronic Components and Devices Research Unit, CDTA, 16081 Alger, Algeria

Abstract. The demand for high-speed networks is increasing in telecommunications and portable devices. In this work, an RF MEMS switch is designed to be integrated into reconfigurable antennas for 5G applications in the 1–6 GHz frequency range. The proposed electrostatic RF MEMS switch requires a very low operating voltage of 3.1 V. In addition, the switch achieves an insertion loss of −0.0085 to −0.0118 dB and an isolation of −46.2 to −31.5 dB in the frequency range 1–6 GHz.

Keywords: MEMS · Reconfigurable antenna · RF MEMS switch · 5G

1 Introduction

The global market is witnessing major advances in high-speed communications and increasing complexities in multiple devices connectivity, especially with the deployment of Fifth Generation (5G) network communications. Today's applications are inherently dependent on real-time multiple device connectivity such as the automotive industry, smart cities, and digital health [1].

As 5G circuits are more demanding in terms of speed, operating frequencies and electromagnetic performance, it requires ideal Radio Frequency (RF) switches. These RF switches should have a near ideal switch performance (Short Circuit & Open Circuit states), high reliability, high switching speed, and low operating voltage to be compatible with mobile applications [2].

Traditional RF switches, such as p-i-n diode and Field Effect Transistor (FET), show relatively high losses and nonlinearities. Moreover, RF Micro-Electro-Mechanical-Systems (MEMS) switches present several clear advantages power consumption, switching speed, and RF performance [2]. There are many possibilities with MEMS switches. The switch can be connected to the signal line in series or shunt with capacitive or contact coupling. This significantly change the switch's mechanical and electromagnetic properties. Different actuation mechanisms are also possible such as electrostatic,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 708–713, 2023.
https://doi.org/10.1007/978-3-031-16281-7_67

magnetic, thermal, piezoelectric actuation, however electrostatic actuation is most dominant due to its near-zero power consumption, high switching speed, small size and high compatibility with integrated circuits fabrication processes [3].

The switch developed in this work is an RF MEMS switch. This switch is designed to be integrated into Reconfigurable Antennas (RA). Namely a microstrip patch antenna that is an optimal choice to design a 5G compatible portable antenna, with an operating frequency range of 1–6 GHz.

2 The Proposed Design

2.1 Structural and Principle of Operation

The proposed electrostatic RF MEMS switch consists of a series-connected metal-contact fixed-fixed membrane, suspended over a microstrip line, with two bottom electrodes. The suspended structure consists of two sets of meanders, two square plates for pull-down, two insulator pieces, and a contact plate as shown in Fig. 1. Large areas have holes for better release and to reduce damping. The switch is entirely metallic, except for the two insulator pieces are made of Silicon Nitride. To prevent a short circuit, another dielectric layer is added on top of the bottom electrodes.

The switch is designed to be normally at the OFF state. The gap in the microstrip line prevents the RF signal from passing through the switch and no voltage bias is applied in this state. To switch to the ON state, a voltage bias is applied to the bottom electrode while the two sides of the switch are grounded. The electrostatic force generated between the bottom and upper plates pulls the switch down. This forces the contact area of the switch to touch both sides of the signal line and connect them, while the two insulator pieces prevent the bias voltage from interfering with the signal. When the bias voltage is removed, the switch returns to its original position.

2.2 Materials and Dimensions

The switch is suspended 1.5 μm over the 20 μm gap in the signal line. The suspended part is entirely made of Gold, except for the dielectric parts, which are made of silicon nitride. This adds to the simplicity of the design as it only consists of two materials. In

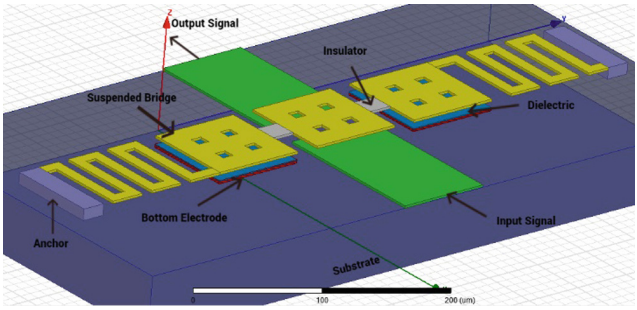


Fig. 1. 3D View of the RF MEMS switch on HFSS

addition, a 0.7 μm layer of silicon nitride is added onto the bottom electrodes. All the dimensions and materials of the switch are summarized in Table 1.

3 Results

To fully characterize the RF MEMS switch, it needs first to be modelled mechanically by simulating the stress response and mechanical switching time. The mechanical stimulation of the switch is done using COMSOL Multi-physics 5.3a Finite Element Modeling (FEM). Electrostatics physics is used in this simulation. The switch's materials are imported from COMSOL's material library, except of air that is added manually. Regarding the electrical setup, DC voltage sources are added to the switch sides and the bottom electrodes. Charge conservation is applied to the air box surrounding the switch needed for electrostatic forces. The studies used in the simulation are: stationary study to get the operating voltage and time dependent study to get the switching speed.

Table 1. Main switch design specifications

Description	Material	Length (in μm)	Width (in μm)	Thickness (in μm)
Contact	Gold	80	80	2
Bottom and Upper electrodes	Gold	80	80	2
Meanders	Gold	80	8	2
Insulator	Si3N4	20	20	2
Half signal line	Gold	140	80	1

Finally, surface graph is used to visualize the stress distribution on the switch as seen in Fig. 2. However, a point at the center of the contact plate is used to measure the vertical displacement and switching time of the switch (Figs. 3 and 4). Then, the electromagnetic modeling was done by studying the RF scattering parameters of the switch using HFSS 15.0.

3.1 Mechanical Modeling

Using the electromechanical module in COMSOL, the bridge is set to be suspended 1.5 μm above the electrode and fixed from both ends (Fig. 1). To switch to ON state a voltage difference is applied between the switch and the electrodes resulting in a downward deflection. A maximum stress level is shown to be 6.6 MPa (Fig. 2). A voltage sweep is done to achieve pull-in and close the 1.5 μm gap. The switch reaches the pull-in state when a voltage of 3.1 V is applied, as shown in Fig. 3. Then, a temporal analysis was made to test the operating speed of the switch. As shown in Fig. 4, the switching time is found to be 64 μs.

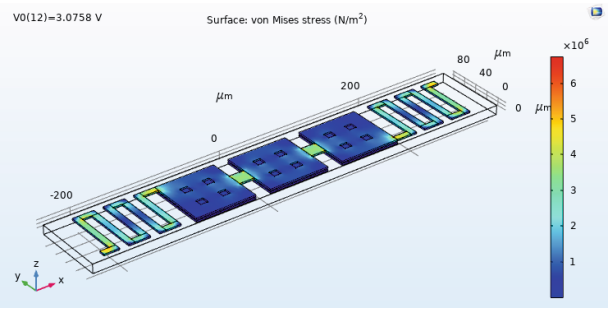


Fig. 2. Stress distribution over the structure

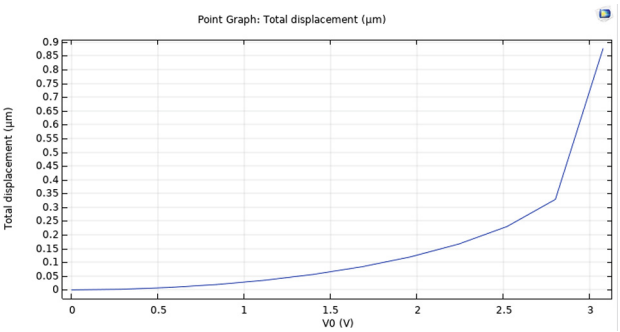


Fig. 3. Total displacement of the contact vs the applied voltage

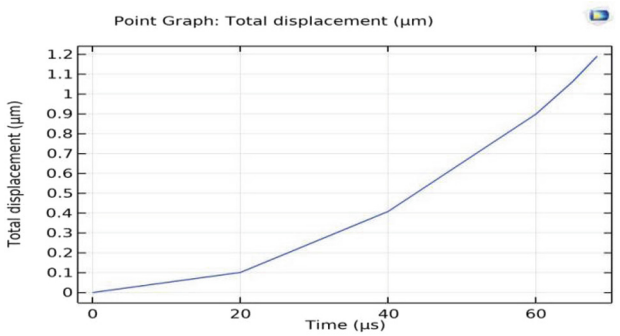


Fig. 4. Switching time of the proposed design

3.2 Electromechanical Modelling

The switch is simulated at two positions. In the OFF state, the bridge is suspended above the signal line; the isolation is then measured from the s-parameters as S21 (Fig. 5). In the ON state, the bridge is down and contacts the two separate parts of the signal line; the insertion losses are then extracted from the s-parameters as S21 (Fig. 6). As shown

in the figures, the switch achieves an insertion loss of −0.0085 to −0.0118 dB and an isolation of −46.2 to −31.5 dB in the frequency range 1–6 GHz.

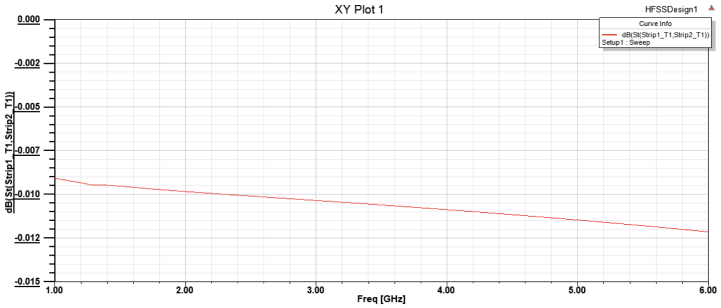


Fig. 5. Insertion loss (S21) from 1 to 10 GHz on HFSS

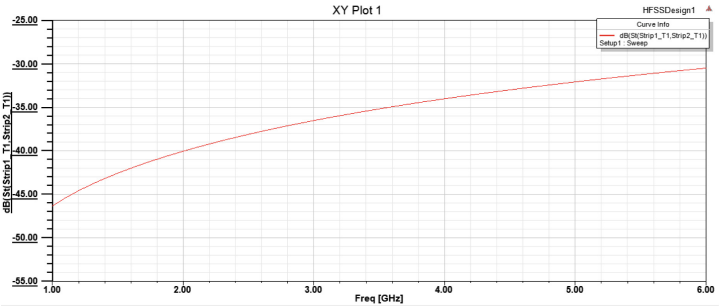


Fig. 6. Isolation (S21) from 1 to 10 GHz on HFSS

4 Discussion

The proposed switch is designed to be series-connected, metal-contact and electrostatically-actuated switch in order to achieve excellent RF performance in the frequency range of 1–6 GHz with very low operating voltage. Table 2 shows comparisons with comparable switches in the literature. It is clear that the operating voltage of the proposed design is much less than the three other switches. Moreover, the RF performance of the proposed switch, represented in insertion and isolation losses, is better than the other switches.

Regarding the Silicon Nitride parts on the top of the switch, they are used to not disturb the signal line with the actuation voltage applied. This proposed design's fabrication is not complex because only two materials are used: Gold and Silicon Nitride. In addition, the dimensions of the switch will facilitate its fabrication as the least thickness needed is 1 μm and the minimum feature is 8 μm.

Table 2. Comparison between the proposed design and previous work

Parameters	Rao et al. [3]	Sathuluri and Sasikala [4]	Thalluri, et al. [5]	Proposed design
Operating voltage (Volts)	7.6	4.5	5.5	3.1
Insertion loss (dB)	−0.05, −0.045 @ 5–40 GHz	−0.55 @ 23 GHz	−0.03 @ 5 GHz	−0.0085, −0.0118 @ 1–6 GHz
Isolation loss (dB)	−38, −10 @ 5–40 GHz	−51 @ 23 GHz	−50 @ 5 GHz	−46.2, −31.5 @ 1–6 GHz

5 Conclusion

In conclusion, a proposed design of an RF MEMS switch was presented. The switch is designed to be integrated into an RA for IoT applications, which operate at the mid-band 5G. Specifically, the frequencies of interest are from 1 to 6 GHz with a bandwidth of about 300 MHz. The switch shows a very low actuation voltage (3.1 V), excellent isolation, and insertion loss (46.2–31.5 dB and 0.0085–0.0118 dB, at 1–6 GHz, respectively).

Finally, weakness points will be identified after the switch's fabrication due to imperfections in materials and fabrication processes. Additionally, further research should be done on other specifications of the RF MEMS switch such as reliability.

References

1. Khana, W., Rehman, M., Zangoti, H., Afzal, M., Armi, N., Salah, K.: Industrial Internet of things: recent advances, enabling technologies and open challenges. Comput. Electr. Eng. **81**(106522) (2019)
2. Rebeiz, G.M.: RF MEMS: Theory, Design, and Technology. Wiley, New York (2004)
3. Rao, K.S., et al.: Design and simulation of fixed–fixed flexure type RF MEMS switch for reconfigurable antenna. Microsyst. Technol. **27**(2), 455–462 (2018). https://doi.org/10.1007/s00542-018-3983-2
4. Sathuluri, M.R., Sasikala, G.: Comprehensive analysis and design of capacitive RF MEMS switches for reconfigurable microstrip patch antenna. Wireless Pers. Commun. **123**, 1–17 (2021). https://doi.org/10.1007/s11277-021-09154-z
5. Thalluri, L.N., et al.: Perforated serpentine membrane with AlN as dielectric material shunt capacitive RF MEMS switch fabrication and characterization. Microsyst. Technol. **26**(6), 2029–2041 (2020). https://doi.org/10.1007/s00542-020-04755-3

2T1M Neuromorphic Synapse with Pt-Hf-Ti Memristor Model

Heba Allah Gamal[1], Hassan Mostafa[2(✉)], and Ayman Haggag[1]

[1] Electronics Technology Department, Faculty of Technology and Education, Helwan University, Cairo, Egypt
heba.gamal@techedu.helwan.edu.eg
[2] Electronics and Communications Department, Cairo University, Cairo, Egypt
hmostafa@uwaterloo.ca, hmostafa@staff.cu.edu.eg

Abstract. Neuromorphic computing systems are the future of the technological revolution due to their containment of multiple processing layers, low energy consumption, high performance, adaptability, faster data processing and ability to work in parallel. Neuromorphic computing mimics the way information is processed inside the human brain compared to Von Neumann architecture. Von Neumann architecture has several disadvantages like high power consumption, slow speed, low performance, and a lower density of information to be processed. Because of above arguments, neuromorphic computing has become the new architecture for processing. Neural computing is the mainstay of any type of Artificial Neural Networks (ANNs) as it processes information in the same way that biological neural systems such as the brain. This is what makes it unique from any other system. ANNs consists of two main parts, neurons and synapses, just like biological neural networks in the human brain. In the past decade, ANNs depended on transistors in their manufacture, but after the discovery of the memristor and its advantages such as nanoscale, low energy consumption, low emitted heat, non-volatile memory, high performance and high storage space, it became a good alternative to the transistor and this will increase effectiveness of Von Neumann structure. In this paper, an electrical synapse circuit consisting of two transistors of MOSFET type and a Pt-Hf-Ti memristor (2T1M synapse) is used. This work is done on Cadence virtuoso simulation environment using Verilog-A VTEAM model for memristor modeling. This paper shows practical results of using Pt-Hf-Ti memristor in 2T1M neuromorphic synapse, as the Verilog-A VTEAM model is based on experimental data rather than other literature results based on simple simulation model. The simulation results reveal that, symmetric read signals used before in literature can't be applied to the Pt-Hf-Ti memristor as the device is asymmetric by its nature. However, a small pulse duration read signal can be applied without causing destruction to the ANN weight stored in the memristor.

Keywords: Neuromorphic computing · Pt-Hf-Ti memristor · Neuromorphic Synapse · Hardware acceleration

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 714–726, 2023.
https://doi.org/10.1007/978-3-031-16281-7_68

1 Introduction

The development of Artificial Neural Networks (ANNs) of various types and their ability to deal with huge amounts of diverse data and perform difficult and complex tasks in many areas through interaction and learning like humans, has become the reason for the tremendous progress in which we live now.

These networks contain commonalities between characteristics and advantages of each of the human brain and computer software. Electronic components are used to represent the nervous system of the human brain like neurons and synapses and in storing huge data through the use of software and algorithms [1]. ANNs are characterized by small size, low power consumption, high speed and high performance in data processing. Parallel method is used when transferring data from memory to processor and vice versa compared to the old Von Neumann processors. This can be achieved largely by using emerging devices that combine processing and memory in the same device such as memristor [10].

The main building block of any ANN is the Neuromorphic Computing block. It contains two main parts: Neuron and Synapse, just like the human brain [1]. This paper focus on synapse that based on memristor as computing and memory unit.

Neurons are the main component of the brain structure and are responsible for remembering and processing information. They are connected to each other through synapses, and the synapse is able to transmit signals between the two neurons it is connected to. The neuron takes the required action according to the signal from the other neuron [1].

There are different types of Artificial Neural Networks (ANNs), depending on the application for which they are used such as Deep Neural Network (DNN) used in computing smart systems, Convolutional Neural Network (CNN) used for pattern recognition and classification, Recurrent Neural Network (RNN) used for handling sequential data like translation and converting audio to text, and Spiking Neural Network (SNN). For example, SNN is the type that most resemble human neural activity in terms of time difference between neurons. This is what makes them, when using a memristor, very similar to a biological neural network [6–9]. Also, these networks contain many layers containing a large number of neurons and synapses used to increase the accuracy in processing information. So, when applying the pro-posed synapse to these networks, it may achieve more accuracy and effectiveness, especially when using a Pt-Hf-Ti memristor.

The Memristor is widely used in synapses [3]. A memristor is an inactive element that has two ends, its resistance changes when a voltage is applied to it, and in the event of this voltage being interrupted, it retains its value, making it work as if it were a non-volatile memory (storage element) [4]. This is what makes it distinct from other elements in addition to its low heat emission, nanoscale size and low energy consumption. This is what made it suitable for use as an electrical synapse.

Every synapse contains a weight that may be positive or negative depending on its value. The greater the weight, the stronger the connection between neurons and vice versa. This is what happens naturally with biological neurons in the brain for learning and training. This weight is multiplied by the neuron's input and this weight is symbolized by the symbol (W). In this case, the memristor stores the weight and updates it according to the training algorithm used [10].

In many literatures, the memristor is used as a synapse, however, it is necessary to take into account the ANN used and its application. Authors in [11] used one crossbar array that has suffers from sneak path and can't handle negative weights. Authors in [12] used two crossbars which could handle negative weights but twice as large as 1M. Authors in [13] used two memristors and one transistor (2M1R). This worked well with SNNs which used one memristor for control and another for storage and a resistance to maintain the weight value to correct any errors that occur or biases applied to the synapses. Authors in [10] used one memristor to store the weight and two transistors to control the reading and writing for machine learning, and is called (synapses 2T1M). This paper will focus only on synapse 2T1M, and future work will be on other types.

The work done in literature with 2T1M neuromorphic synapse [2, 10] was based on titanium memristor and Linear Ion Drift memristor model [4], while in this literature [18] used spintronic memristor and used the same Linear Ion Drift model. The Linear Ion Drift model is a very simple model with lowest accuracy and doesn't describe the real memristor behavior [5]. So, in this paper use another model that is able to show all the characteristics and advantages of the memristor and also used a type of memristor capable of being used within ANNs with high accuracy and efficiency. In this paper, the 2T1M neuromorphic synapse is investigated by using VTEAM Verilog-A model [14] that is fitted to experimental data from Pt-Hf-Ti memristor [15]. Due to the asymmetric I-V characteristics of the Pt-Hf-Ti memristor [15], asymmetric read signal is applied rather than symmetric one [2], as illustrated in Sect. 4.

The rest of the paper is organized as follows, the Pt-Hf-Ti memristor background is introduced in Sect. 2. The 2T1M synapse used in [10] for applying online gradient descent algorithm in Multilayer Neural Networks (MNNs) is introduced in Sect. 3. The simulation results of using VTEAM Verilog-A model for modeling Pt-Hf-Ti memristor in the neuromorphic synapse are introduced in Sect. 4. Finally, the paper is concluded in Sect. 5.

2 Pt-Hf-Ti Memristor

The Pt-Hf-Ti memristor is one of the types of memristors based on metal oxide materials, which are called anionic devices and depend on the movement of oxygen gaps when an electric field is applied. It is similar in its structure to the installation of a capacitor containing a metal insulator (MIM) as shown in Fig. 1. It consists of three layers: platinum (Pt) and titanium (Ti) and they represent the electrodes (metal materials) and in the middle is the insulating material, which is hafnium dioxide (HfO2). These materials have the ability to achieve the physical and chemical properties that enable them to reach the nanoscale. The middle layer is called the active region of memristor, which works to switch the resistance of the low and high to memristor which switching from ON state to OFF state by an external electric field applied to the memristor [16].

It's working principle depends on switching the resistance through ionic transport through oxygen gaps in the active region (HfO_2). When a positive voltage is applied by the external electric field to the electrode (Pt), the oxygen gaps are transferred to the electrode (Ti) and there is a low resistance and vice versa. The change in the case of resistance depends only on the magnitude of the applied voltage and not the polarity.

Possibility of adding other layers of insulating material in Pt-Hf-Ti memristor to improve the switching speed compared to the performance of a single layer [16].

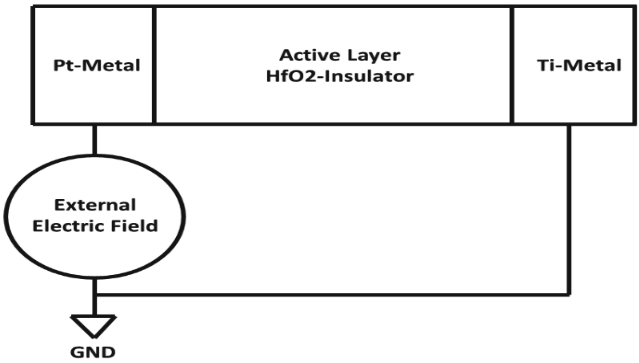


Fig. 1. Pt-Hf-Ti memristor structure with external electric field connections to change the internal resistive state.

It is preferred to use the HfO2 memristor in this paper because of its simple structure, high resistance switching stability, high switching speed, high reliability, low power consumption, and compatibility with CMOS technology [17]. It can be designed as a crossbar to allow memory and computation integration and a parallel computing procedure for the weight matrix used to collect the input signals and the weights of synapses. Thus, it will suit multi-layer neural networks, and in the applications of logical computing, neural computing, non-volatile memories, analog and digital circuits.

3 2T1M Neuromorphic Synapse

The 2T1M synapse schematic diagram is shown in Fig. 2 used in scalable online learning algorithm in neural networks in [10], with titanium memristor replaced by Pt-Hf-Ti memristor. Synapse consists of a one Pt-Hf-Ti memristor that stores the weight (W_t). This weight shows how important the connection between the two neurons is and the strength of the synapse and it is updated according to the training algorithm used. In order for the ANN to function efficiently, it needs three operations: reading, writing, and training. So, the transistors control the reading and writing processes (storing the stored weight) according to the signal input to it. Based on the reading and writing processes, the network is trained and its tasks are implemented. High conduction transistors of MOSFET type (N-Channel - P-Channel) were used so that the input voltages of the transistors are very close to the input voltage of the Pt-Hf-Ti memristor (i.e., the ON resistance of the transistors is very small) and the control signal (e) which is the input signal (aX) is of high and suitable voltage. To achieve this, a scaling factor (a) is used so as not to change its value when using it with Pt-Hf-Ti memristor and make it more linear, is used. The MOSFET N-Channel transistor is connected to the positive voltage while the MOSFET P-Channel is connected to the negative voltage as shown in Fig. 2 so that the transistor is in the deep triode region.

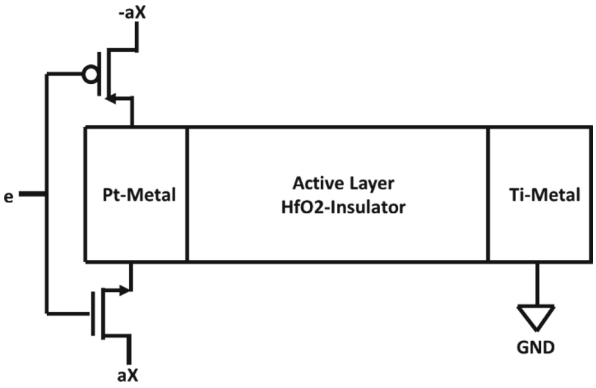


Fig. 2. 2T1M Synapse schematic in [10] used Pt-Hf-Ti memristor.

It is known that any change in the parameters of the memristor will change the value of the memrisance. When training the network, a difference may result in the network output as a result of entering the data through it necessary for training, and then the weight is updated accordingly. The network uses control signals, these signals are square wave for specific time periods that depend on the error in order to update the weight according to the shape of the signal and its time duration.

The original signal used in [10] is a symmetric square signal. This means that the positive and negative halves of the reading and writing stages are identical with time, and this prevents accumulating mistakes during the repeating reading and writing stages. The time period for the reading stage is (40 ms), while the time period for the writing stage is (60 ms), and the total time period of the control signal is (100 ms), as shown in Fig. 3(a). This signal is effective during simulation, training and updating weights. This is useful when applied to ANNs and machine learning tasks, but an error appears after few periods of time operation based on Eq. (2) [2, 10]. The reading stage involve using voltage signals proportional in time with the polarity of the voltage opposite to it, so there may be a destructive reading of the stored state in synapse. While the writing stage is determined by the error (y) and converted into a time signal (error appears during writing), which shows the result of the feedback of the algorithm used [10] between the required output (d) and the actual output (r).

$$y = d - r \tag{1}$$

$$r = WT.X \tag{2}$$

This signal cannot be used, because practically it can't be achieved. There is no signal generator that can give an ideal signal with equal positive and negative halves in the pulse width.

It is necessary to pay attention to the reading stage so that it doesn't distract the state of memristor and the error that appears after a period of reading cycles. So, in order for the synapse to work in practice, an asymmetric signal with two unequal positive and negative halves in the pulse width of the reading stage is applied. The mismatch

(α) causes a destructive reading (destroying the stored weight) after longer operating operation cycles, which may also cause a write error as a result of the accumulated weight error appearing after long periods of operation with a value greater than or equal to the previous stored weight. As a result of the appearance of wrong weights, this will greatly affect the machine learning tasks of ANNs.

In [2], two signals with the same voltage and the same period values of the reading and writing stages were used, but in the reading stage, the positive and negative halves aren't equal, as shown in Fig. 3 (b). The time period for the reading stage is (40 ms) while the time period for the writing stage is (60 ms) and total time period of the control signal is (100 ms). Also in Fig. 3 (c), the same time periods are used. However, as a pulse width in nanosecond is used instead of millisecond, this may have a slight change in memrisance in the reading stage. The nanosecond scale has a cumulative effect of the stored weight after more operating periods than the millisecond scale. When applying those signals in Fig. 3 (b) and (c), the error that appeared after hundreds of time periods compared to use the original signal in shown Fig. 3 (a) [10].

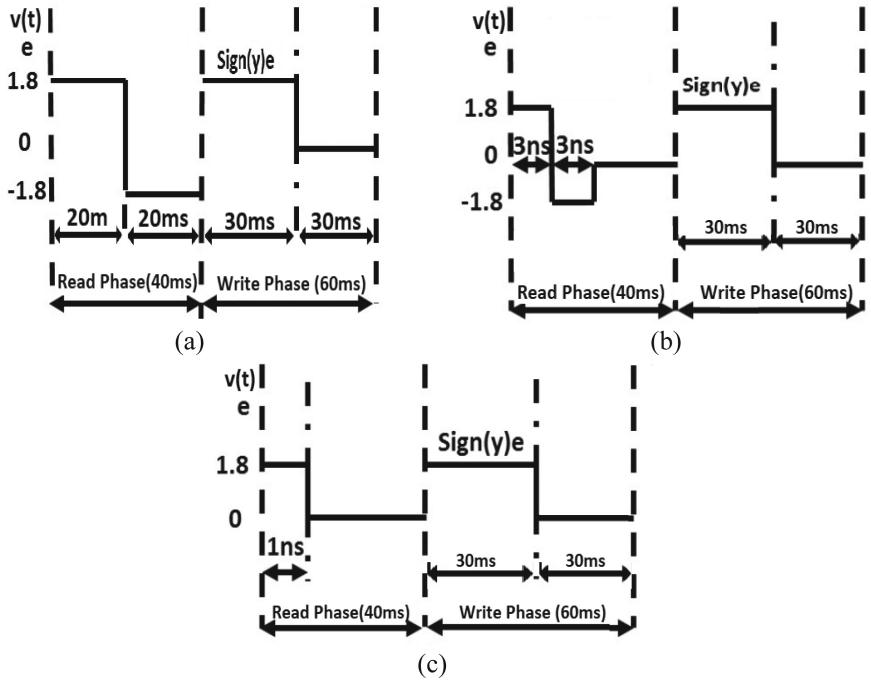


Fig. 3. A full period (100 ms) of the control signals in other literature, (a) the original control signals in [10], while (a) and (b) the suggested control signals in [2].

While authors in [18] used the same signals with the same voltage but with different values of reading and writing periods, and the reading stage, in which positive and negative halves aren't equal, and the pulse width is nanosecond instead of millisecond, achieved a less destructive reading compared to the signal used in [10] and shown in

Fig. 3 (a). The nanoscale of signal has the benefit of making the memristor more linear, works better and has fewer errors.

The literature [2, 10, 18] used the Linear Ion Drift Memristor model with different types of memristor, but this model is preliminary and impractical, however, in this paper a working VTEM model was used instead.

4 Simulation Verification

The memristor Verilog-A VTEAM model in [5] is used in Cadence Virtuoso. The CMOS 130 nm technology is used for both transistors and the parameters of the components are shown in Table 1.

The simulation verification of the Pt-Hf-Ti memristor in 2T1M neuromorphic synapse is done in two steps. Firstly, the I-V characteristic verification of fitting VTEAM model to the Pt-Hf-Ti memristor to make it close to linear is done as in [14, 15]. Secondly, applying suitable control signals (e) to the neuromorphic synapse based on the first verification is done.

Table 1. Circuit parameters.

	Parameter	Value	Description
Transistors	W	2 um	Width
	L	130 nm	Length
Input	A	1 mV	Scaling factor
	X	10	Machine learning input
Control signal	e	±10	Control signal amplitude
Timing	T	2 us	Period
	Tread	1 us	Reading period
	Twrite	1 us	Writing period
Memristor	R_{on}	50	Minimum resistance
	R_{off}	1000	Maximum resistance
	D	3 nm	Width
	v_on	−0.002	Threshold voltage in VTEM
	v_off	0.009	Threshold voltage in VTEM

4.1 I-V Characteristics of Pt-Hf-Ti Memristor

The VTEAM general model can be used in a wide range of applications because it is flexible, simple, accurate and with high computational efficiency. It depends in its work on the threshold voltage which results in a more practical and less error model than other

models. So, it is preferred to be use, however, one of its disadvantages is that it is linear at high frequencies only.

The I-V characteristics of the Pt-Hf-Ti memristor when applying 1 V sine wave at different frequencies (1 MHz, 50 MHz, 100 MHz, 1 GHz) is shown in Fig. 4. As can be seen in Fig. 4, the higher the operating frequency of the Pt-Hf-Ti memristor, the more linear is the I-V characteristic is, as it should be for the memristor [14, 15].

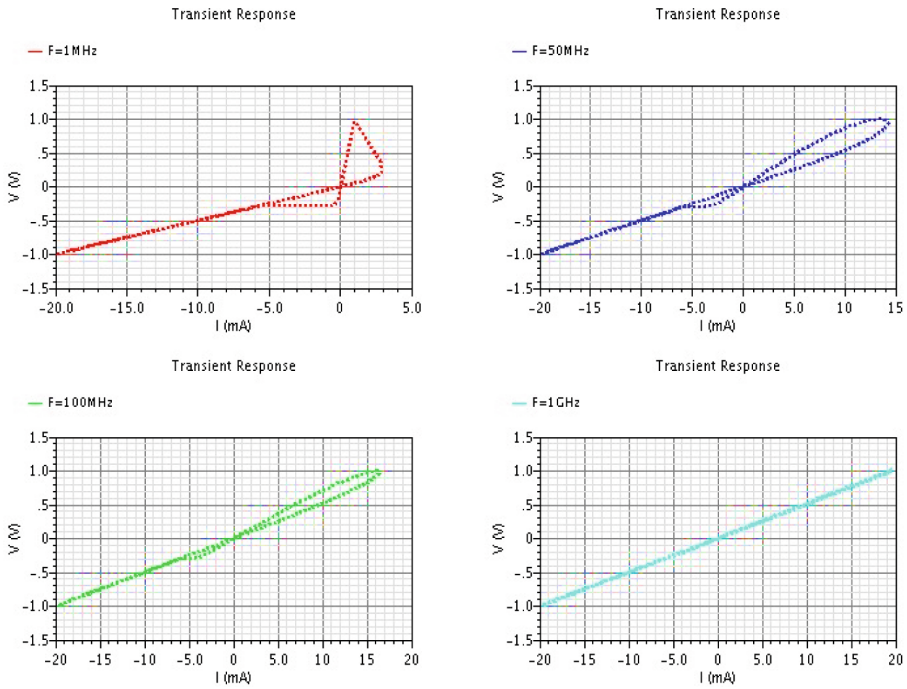


Fig. 4. The I-V characteristic curve for the Pt-Hf-Ti memristor with applied 1 V sine wave at frequency = 1 MHz, 50 MHz, 100 MHz and 1 GHz.

4.2 Control Signals Study

It is observable from Fig. 4 that the memristor I-V characteristic curve is asymmetric except at very high frequencies (from 1 GHz and above). This results in the restriction of applying a symmetric read, like the two signals used in [2] and also when applying the original signal used in [10] because they work at low frequencies (25 Hz), and also when applying what is suggested by the literature [18].

Hence, in this paper, in order for the Pt-Hf-Ti memristor to work well through the VTEAM model, three control signals were proposed to compensate for the destruction of the stored weight after several operating periods. These signals are inspired by the original signal and also used by the literature [2] in Fig. 3 (c). These signals are preferred

because they fit the VTEAM model and work with it with high efficiency and with nano time periods. Each of them differs from the others in the pulse width of the reading stage.

Figure 5 shows the three control signals used in the two processes of reading and writing the weight of the synapse (the input voltage of the transistors). It contains two stages, the first stage is for reading the weight, while the second stage is for writing the weight (updating the weight). Then the time period for reading and writing becomes (2 us), the reading period is reduced to (1 us) instead of (40 ms) and the period of writing is reduced to (1 us) instead of (60 ms). A changing pulse width in the reading stage is used in order to reduce errors at a suitable frequency to make the Pt-Hf-Ti memristor more linear with changing threshold voltages values for the memristor, as shown in the Table 1.

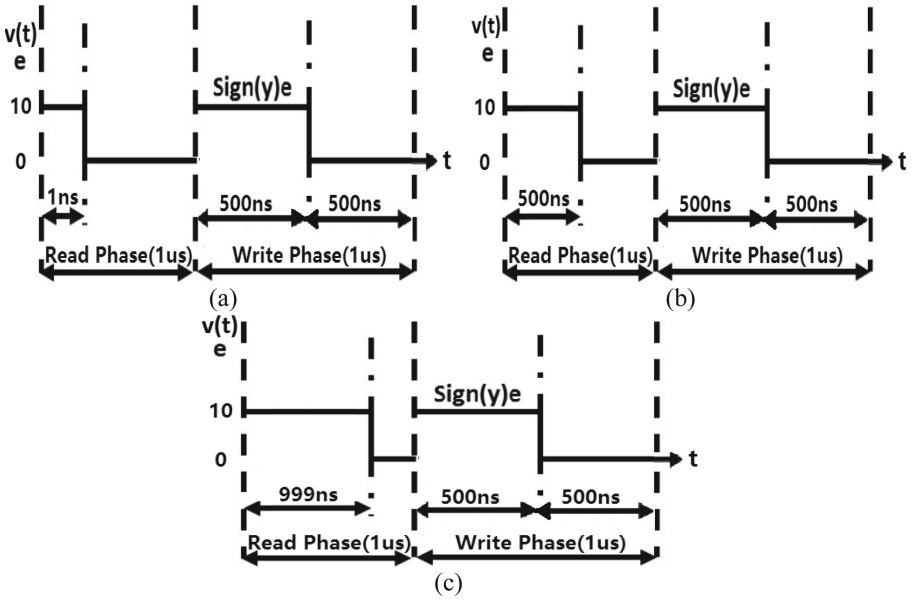
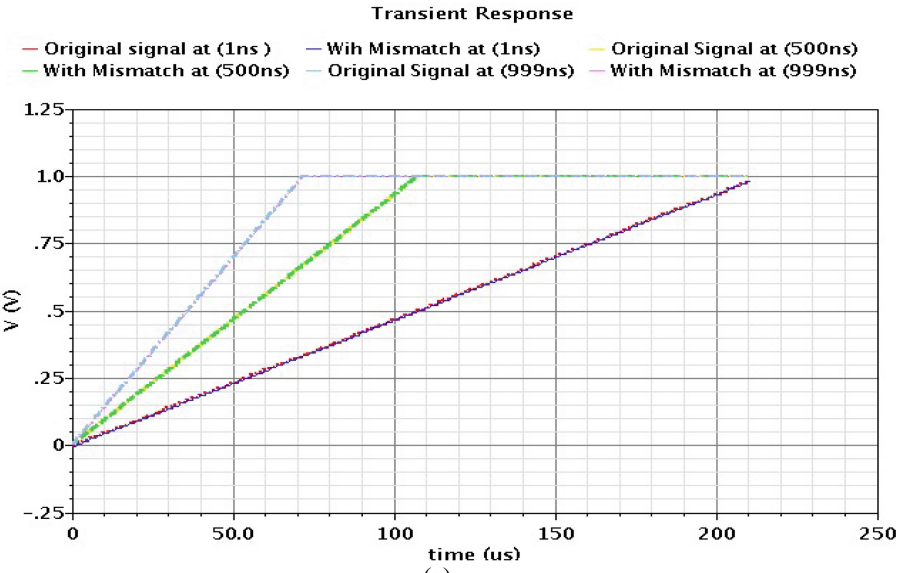


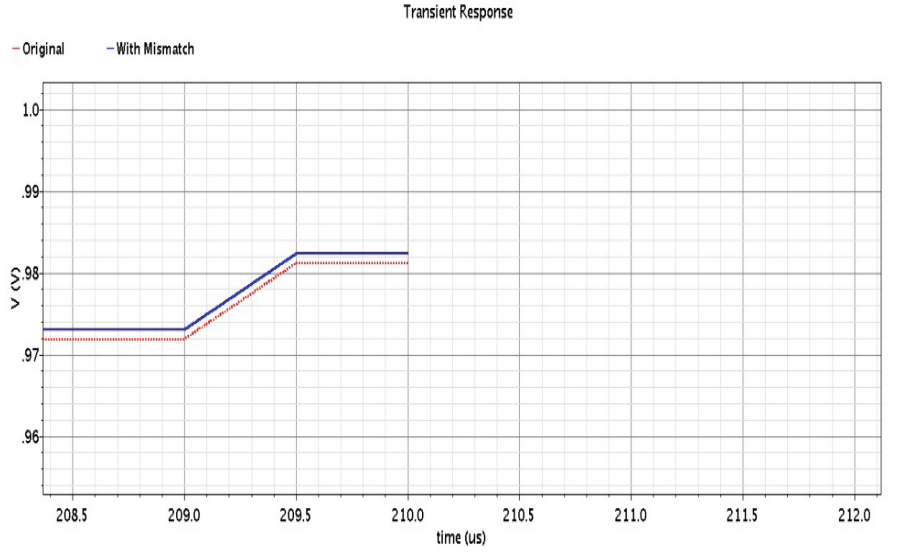
Fig. 5. A full period (2 us) of the modified three control signals e with pulse width mismatch alternative to the signal in [2], for both read and write of synapse weight. (a) is signal A, (b) is signal B and (c) is signal C.

As can be seen in Fig. 5, it is possible to enlargement the pulse width for the reading stage from (1 ns) to (999 ns). However, it is not possible to make the signal in picosecond unit, because the maximum signal that can be used is nanosecond, this makes Pt-Hf-Ti memristor linear.

Figure 6 shows use of the proposed control signals in Fig. 5, where they were also applied in case of mismatch. When applying the three signals we still get destructive readers, however, the stored weight will be destroyed after long periods (500 ps). This makes it distinct from [2] which destroys the weight after (310 ps). This is due to the fact that the duration of pulse is very small and only small changes occur to the state variable of the memristor.

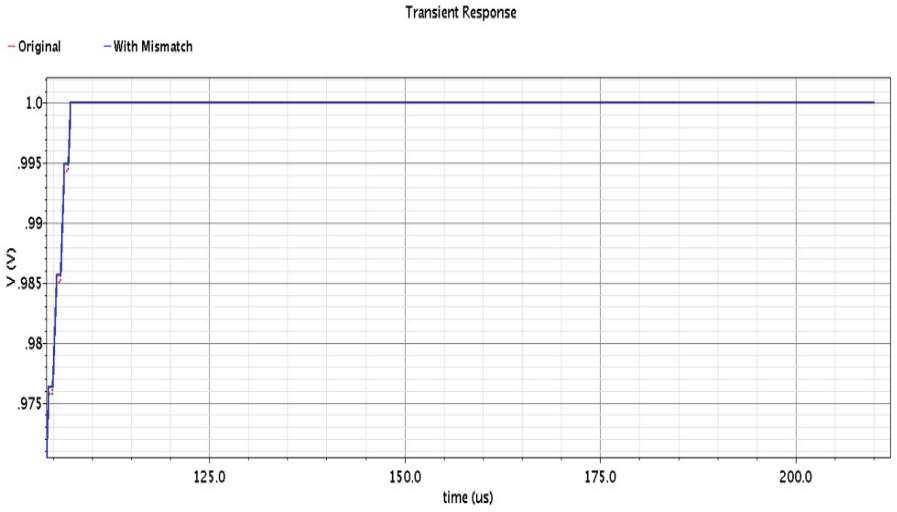


(a)

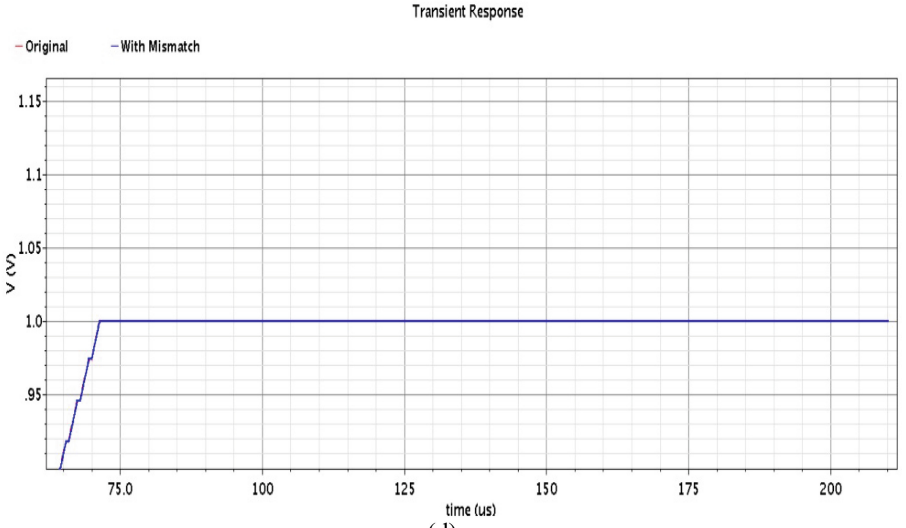


(b)

Fig. 6. (a) is simulation result for the three modified signals. (b) is zoom of the signal (1 ns) for applied of the modified signal A, (c) is zoom of the signal (500 ns) for applied of the modified signal B, (d) is zoom of the signal (999 ns) for applied of the modified signal C.



(c)



(d)

Fig. 6. continued

As shown in Fig. 6(b), it is found that the state variable changes slightly in the case of signal with and without the mismatch. While Fig. 6(c), It is also found that the state variable changes in a small way. While Fig. 6(d), The simplest is in the case of signal with and without the mismatch, it is found that the state variable almost has the same effect in the case of signal with and without the mismatch.

The suggested signals are applied to the synapse, i.e., in the case of mismatch and without it, for the reading stage to compare the condition of the synapse during operation. It is found that the higher the pulse width the more destructive reading occurred. Meaning

that the signal (a) is less destructive than (b) and (c). The error occurred after (500 ps) of the operating cycles due to the mismatch of the reading stage. The Pt-Hf-Ti memristor proved that it is working fast at high frequencies with less energy consumption with the ability to work at nanosecond scale.

5 Conclusion

Neural computing systems are the future of technological revolution, and therefore they need continuous development. They provide a new architecture for processing, including artificial neural networks (ANNs) because they process information in the same way as biological neural systems. Memristor is a newly discovered element by Leon Chua with a mechanism of action similar to the biological learning mechanisms. It has the ability to work as a non-volatile memory, in addition to its small nano-size, low energy consumption, low heat emitted, high performance and high storage space, which made it suitable for simulating behavior of synapse and dealing with different updates of weights and high performance of ANNs. A method is proposed to perform the weight update without causing damage to the stored weight after a number of operating periods by constructing a neural structure that represents an electrical synapse. Our proposed neural structure is a small circuit containing a Pt-Hf-Ti memristor to store synaptic weight and two CMOS transistors to control the circuit. It can also be used with multi-layer Artificial Neural Networks (MNNs) without destroying the weight until after very long time periods.

Acknowledgement. This work was partially funded by ONE Lab at Zewail City of Science and Technology and Cairo University, NTRA, ITIDA, ASRT, Mentor Graphics, NSERC.

References

1. Basheer, I.A., Hajmeer, M.: Artificial neural networks: fundamentals, computing, design, and application. J. Microbiol. Methods **43**(1), 3–31 (2000)
2. Hassan, A.H., Mostafa, H.: Nondestructive reading and refreshment circuit for memristor-based neuromorphic synapse. In: 2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS), Dallas, TX, USA, pp. 642–645 (2019)
3. Krestinskaya, O., James, A.P., Chua, L.O.: Neuromemristive circuits for edge computing: a review. IEEE Trans. Neural Netw. Learn. Syst. **31**(1), 4–23 (2020)
4. Strukov, D.B., Snider, G.S., Stewart, D.R., Williams, R.S.: The missing memristor found. Nature **453**(7191), 80–83 (2008)
5. Kvatinsky, S., Talisveyberg, K., Fliter, D., Kolodny, A., Weiser, U.C., Friedman, E.G.: Models of memristors for SPICE simulations. In: 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel, Eilat, pp. 1–5 (2012)
6. Aloysius, N., Geetha, M.: A review on deep convolutional neural networks. In: 2017 International Conference on Communication and Signal Processing (ICCSP) (2017)
7. Gu, J., et al.: Recent advances in convolutional neural networks. Pattern Recogn. **77**, 354–377 (2018)
8. Pfeiffer, M., Pfeil, T.: Deep learning with spiking neurons: opportunities and challenges. Front. Neurosci. **12**(774), 1–18 (2018)

9. Shrestha, A., Mahmood, A.: Review of deep learning algorithms and architectures. IEEE Access **7**, 53040–53065 (2019)
10. Soudry, D., Di Castro, D., Gal, A., Kolodny, A., Kvatinsky, S.: Memristor-based multilayer neural networks with online gradient descent training. IEEE Trans. Neural Netw. Learn. Syst. **26**(10), 2408–2421 (2006)
11. Zhang, Y., Wang, X., Friedman, E.G.: Memristor-based circuit design for multilayer neural networks. IEEE Trans. Circuits Syst. I Regul. Pap. **65**(2), 677–686 (2018)
12. Prezioso, M., Merrikh-Bayat, F., Hoskins, B., Adam, G., Likharev, K.K., Strukov, D.B.: Training and operation of an integrated neuromorphic network based on metal-oxide memristors. Nature **521**(7550), 61–64 (2015)
13. Hu, M., Chen, Y., Yang, J.J., Wang, Y., Li, H.H.: A compact memristor-based dynamic synapse for spiking neural networks. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **36**(8), 1353–1366 (2017)
14. Kvatinsky, S., et al.: VTEAM: a general model for voltage-controlled memristors. IEEE Trans. Circ. Syst. II: Express Briefs **62**(8), 786–790 (2015)
15. Yalon, E., et al.: Resistive switching in probed by a metal–insulator–semiconductor bipolar transistor. Electron Device Lett. **33**(1), 11–13 (2012)
16. Mohammad, B., et al.: State of the art of metal oxide memristor devices. Nanotechnol. Rev. **5**(3), 311–329 (2016)
17. Wu, L., Liu, H., Li, J., Wang, S., Wang, X.: A multi-level memristor based on Al-doped HfO 2 thin film. Nanoscale Res. Lett. **14**(1), 1–7 (2019)
18. Abdelmagid, Y.K., et al.: Investigation of DW Spintronic memristor performance in 2T1M neuromorphic synapse. In: 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES), pp. 573–577. IEEE, October 2020

Embedded Implementation of Signal Pre-processing for Tactile Sensing System

Moustafa Saleh(✉), Yahya Abbass, and Maurizio Valle

University of Genova, Genoa, Italy
moustafa.saleh@edu.unige.it

Abstract. Smart tactile sensing system has been a subject of research in many application domains such as prosthetics and robotics. Embedding signal pre-processing methods (i.e., filters) along with processing algorithms (i.e., machine learning) into miniaturized electronic units enhance the extraction of high-bandwidth information (e.g., slippage detection). However, it is challenging due to the high computational costs and the real time requirements. This paper proposes a lightweight implementation of pre-processing method for multichannel tactile sensing system. We targeted two filtering methods, Finite Impulse Response (FIR) and Exponential Moving Average Filter (EMAF). The paper presents the analysis of the implementation performance on hardware i.e., number of clock cycles, execution time and touch detection accuracy. Experimental results show that EMAF is more effective than FIR when it comes to the hardware complexity. This means that the computational cost for implementing such pre-processing filter is negligible and thus acceptable for time, and hardware constraint tactile sensing system.

Keywords: Tactile sensing system · Signal processing · Filtering methods · Embedded implementation

1 Introduction and Related Work

Tactile sensing in humans is one of the fundamental sensory modalities (visual, auditory etc.) that plays an important role in conveying information to the brain about objects in contact (e.g., contact surface, roughness, shape, grasp stability and slip detection [1] etc.). Considerable scientific efforts have been devoted to enhance the motor control in robotics to carry-out human-like movements such as manipulation and exploration tasks [2]. Performing such tasks successfully requires a tactile sensing system that can extract high bandwidth tactile information.

Signal processing and feature extraction algorithms are crucial to properly decode and translate tactile sensor signals into useable tactile information. Over the past two decades, the usage of tactile sensing systems in practical applications has been limited due to some limitation in the performance of the multi-channel sensing system (i.e., sensor and signal conditioning circuits). For example, signals from massive amount of tactile sensors may be noisy and containing irrelevant information [3]. Thus, this arises

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 727–736, 2023.
https://doi.org/10.1007/978-3-031-16281-7_69

the demand of developing efficient sensing systems at both the hardware level (e.g., highly sensitive sensors and acquisition circuits) and the software level (e.g., signal processing methods). As on the software level, general statistical algorithms and machine learning techniques have been applied to the problem of extracting tactile information. The Fast Fourier Transform (FFT) and Discrete Wavelet Transform (DWT) are leading techniques for preprocessing and simplifying signal output from sensors. The FFT is most used for temporal signals from sensors. DWT is often used when FFT results fail to produce satisfactory features of the touched object (e.g. surface characteristic) [4]. It is also used for feature extraction and its filters can also be applied for the processing of raw signals; DWT can be used for creating high-pass and lowpass filters that can construct the original frequency and the temporal information. However, the deployment of these two techniques for signal filtering result in a high computational load as both methods require the computation of the spectrum and then the spectrum inverse. Therefore, large memory is required, thus the sensing system is characterized with high latency. Statistical and learning algorithms including decision trees (DT), support vector machines (SVM), extreme learning machines (ELM), gradient boosting machines (GBM), maximum likelihood estimations (MLE), k-means clustering (KM), and k nearest neighbors (KNN) are used to categorize and identify many properties of the touched object [5].

Prior performing any of the aforementioned algorithms, signal pre-processing is required, such as removing signal disturbances with appropriate filters. This allows the sensing system to well extract features and interpret the measured information. However, implementing such methods (i.e., filters) along with processing algorithms (i.e., touch modalities classification algorithms) onto the embedded tactile sensing system poses some challenges on system performance in terms of computational cost, response accuracy [6] and time delay. To the extent of our knowledge, such embedded implementation of signal pre-processing and processing methods for a real tactile sensing application is still lacking. Where in most cases data are collected and stored to then be processed offline on PC hosts. Thus, assessing the performance of such processing methods on embedded sensing system hardware is strongly recommended.

This paper proposes a lightweight implementation of pre-processing filtering method for multi-channel tactile sensing system prepared to be used in real application. For this purpose, an end-to-end distributed sensing system was developed on board aiming to collect data corresponding to human-like touches. Two non-recursive digital filters have been deployed on the Interface Electronics (IE) as the acquisition block of the sensing system and applied onto the collected data: FIR and EMAF. These two filters have linear phase characteristics, which unlike other filters (e.g., Infinite Impulse Response and decimation) are more stable and allow retaining the shape of the original signal after filtering. Experimental tests have been carried out to assess the behavior of both filters running on the IE. To select the filter with better performance, analysis metrics have been measured such as Signal-to-Noise Ratio (SNR) of the output tactile signal, detection of the touch events and computational complexity of each technique. Experimental results show that with EMAF few computations are required to filter out tactile signals which unlikely when using FIR. Thus, making EMAF more suitable for a time and hardware constraints tactile sensing system.

This paper is organized as follows: Sect. 2 describes the tactile sensing system components, discusses the proposed pre-processing methods and the implementation methodology onto the IE. Section 3 presents and analyzes the performance of the implemented filters. Finally, Sect. 4 concludes the paper.

2 Methodology

2.1 Tactile Sensing System Architecture

Figure 1 shows the structure of a typical tactile sensing system. The system includes 1) a sensing array, 2) an IE for data acquisition, and 3) an embedded digital processing unit for tactile data decoding. Sections below just introduce the sensor array structure, and the IE blocks that have been involved in the presented work. Where the proposed filtering methods have been implemented into the IE.

Fig. 1. Tactile sensing system block diagram

Sensor Array Structure
Sensing arrays reported in [7] require an additional fabrication step to realize skin patches that can be integrated into different applications. The sensor array is based on P(VDF-TrFE) poly(vinylidene fluoride trifluor-oethylene) piezoelectric polymer sensors. Figure 2 shows the structure of the skin patch for the palm. The sensing array used in this work was prepared and integrated following the same approach described in [8]. The arrays were shielded using conductive tapes (Model tesa 60262) and then protected using a thin flexible protective layer (Art. 5500 Dream, Framisitalia). The structure was

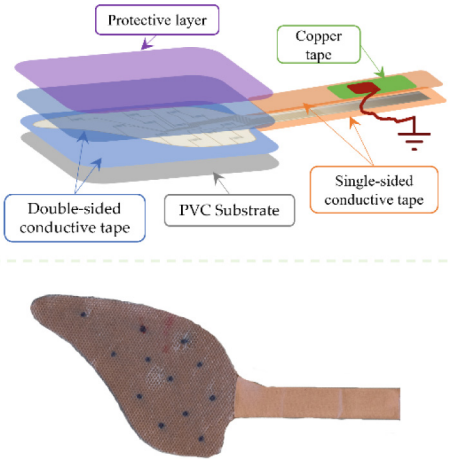


Fig. 2. Top: structure of the skin patch, Bottom: Picture of a real sample of the skin patch.

mounted on a flexible substrate (i.e., PVC). The shielding layers were connected to the ground reference of the embedded electronics using a self-adhesive copper foil tape and a wire.

Interface Electronics Block

Figure 3 depicts the Printed Circuit Board (PCB) of the IE that acts as the acquisition board of the tactile sensing system. It is based on two main off-the-shelf components: BL600 module [9] and DDC232 [10] current-input analog-to-digital converter. The DDC232 converter offers simultaneous sampling for 32 bipolar tactile sensors with configurable sampling rate of up to 6 kHz. The BL600 contains an ultralow-power microcontroller based on an ARM Cortex M0 chip, it is used for retrieving, pre-processing, and transmitting tactile data. The presented design can be powered supply through USB cable, at which is also used for transmitting sensor data throughout the system. The IE has been preliminary validated in previous work [11]. In the present study, the IE was configured to collect and process tactile data from 1 sensor at 2 KSamples/second. It is also endowed with signal pre-processing to detect contact events. Data are transmitted to the PC through a USB cable.

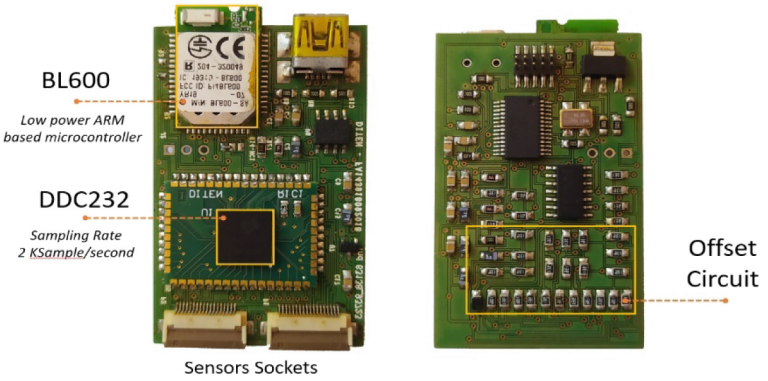


Fig. 3. IE printed circuit board of the tactile sensing system

2.2 Signal Pre-processing Methods Description

In this work we adopted and implemented into the IE two filters as pre-processing methods: FIR and EMAF. The implementation process is introduced and the results of the two methods are discussed in next sections.

Filtering Methods

Finite Impulse Response

FIR is a well-known digital signal processing filter. The filter output y(n) is a result of discrete-time convolution process between input signal x(n) and the impulse response h(k) of the filter (filter coefficients), see Eq. 1. The design of the FIR filter is composed of two main parameters: cutoff frequency and filter order. The filter order generally is

recommended to be high to have more accurate results. Thus, implementing high-order filters to smoothen signals from 32 channels will introduce a huge delay in the extraction of tactile information and thus in the feedback loop. Therefore, a trade-off between the accuracy of the filter and the order should be considered while designing the filter.

$$y(n) = \sum_{k=0}^{N-1} h(k)x(n-k) \tag{1}$$

Exponential Moving Average Filter
Moving average filter (MA) is one of the popular digital filtering techniques that can smoothen all kind of data and reduce random noise in the data. EMAF is a type of MA filter that operates with low computational burden and can be implemented easily and efficiently. EMAF filter computes a weighted average of time ordered sequence by applying to the previous inputs weights that decrease exponentially [12].

The EMAF on input x[n] is expressed as in Eq. 2:

$$y(n) = ax[n] + (1-\alpha)y[n-1] \tag{2}$$

where x[n] is the current input, y[n] is the current output, and y[n − 1] is the previous output; α is the smoothing factor ranges between 0 and 1. As α decreases, high frequencies are attenuated. The design of these filters is introduced in the next sections where experiments have been carried out to define the filters parameters/coefficients.

2.3 Experimental Setup and Filters Implementation

To define the parameters of the filters – for the purpose of implementation – we have arranged an experimental setup (see Fig. 4) to study the response of the sensing system to different touch patterns that might be incorporated while the skin patch interacts with objects. The setup is composed of a skin patch placed on the top of a strain gauge load cell (Tedea Huntleigh, Model 1042) and faced upside. The load cell was used to measure the force applied on its surface where the skin patch is placed. The force stimulus was conditioned by a PXIe-4330 (NI, US) conditioning board while the charge developed by the sensor (response) was conditioned either by the PCB Sensor Signal Conditioner (482C54) or by the IE as illustrated in Fig. 4. A LabVIEW software developed on a National Instruments PXI system was used to collect, visualize, and save the force stimulus and the charge response.

Three touch patterns were selected, Tapping Touch (TT), Press-Hold-Release Touch (PRT), and Continuous Touch (CT). During the tests, the experimenter applied one of the patterns on a single sensor using his/her finger. For example, Fig. 5 and Fig. 6 show the three patterns applied on a single sensor and the corresponding charge response. Figure 5 shows the response of a single sensor to the TT and PRT patterns. Similarly, Fig. 6 shows the CT pattern and the corresponding sensor response.

The signal peaks are arranged in a sequence reflecting the fact that the touches were applied to the sensors sequentially. The PRT pattern was presented by the sensor by two bursts corresponding to the press and release events, while in-between the bursts there was some wiggling.

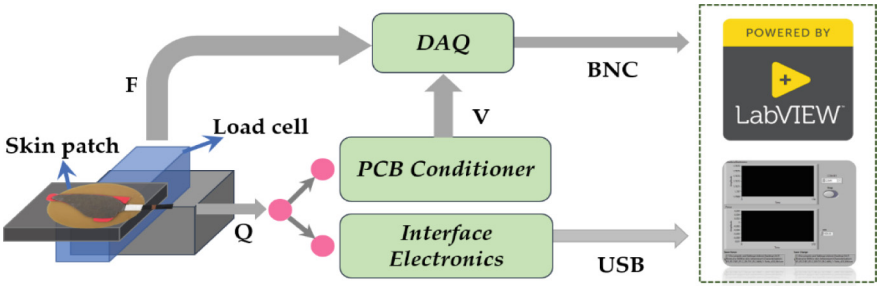


Fig. 4. Experimental setup block diagram

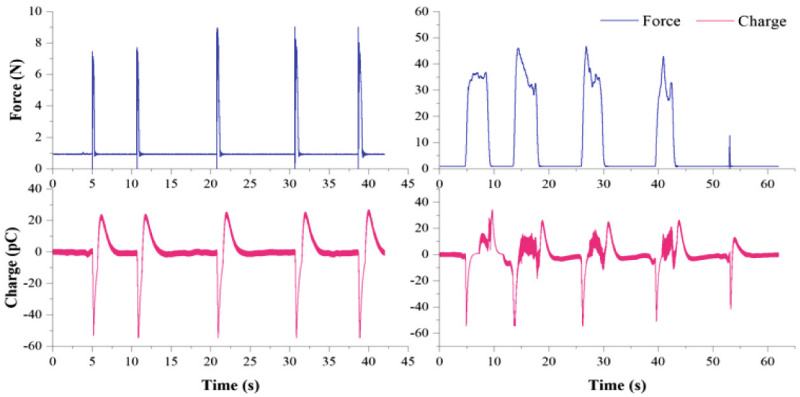


Fig. 5. Response of the single sensor to Tapping Touch (left) and Press-Hold-Release Touch patterns (right)

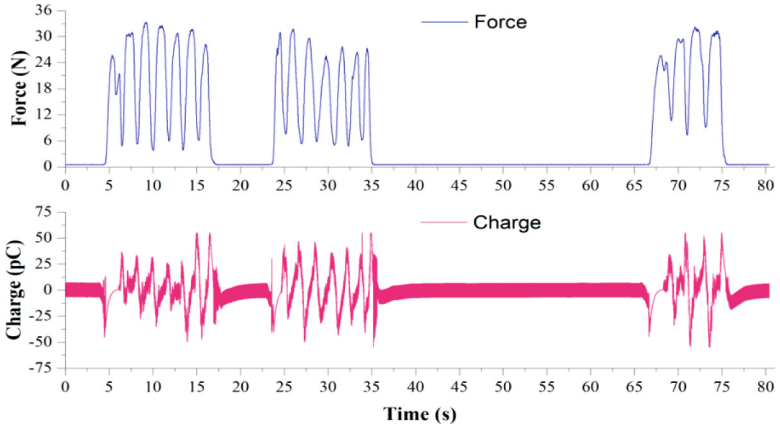


Fig. 6. Response of single sensor to a Continuous Touch pattern

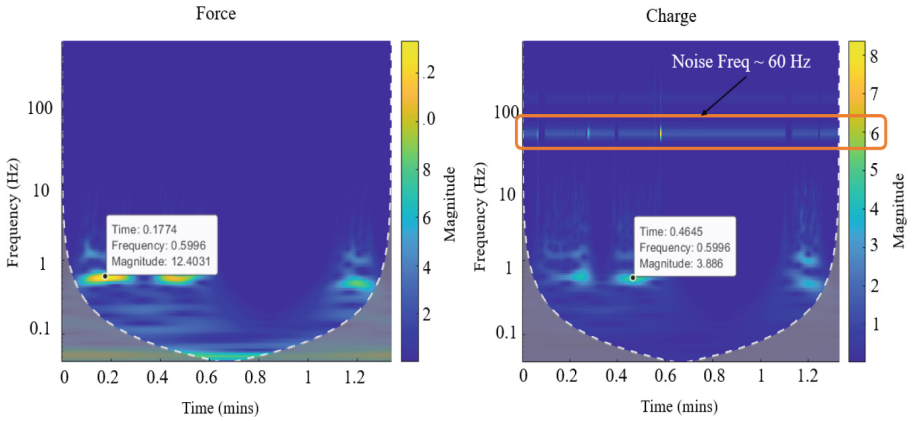


Fig. 7. Continuous wavelets transform (CWT) of the applied stimulus and the sensor response in Fig. 6

Data Collection

To assess the behavior of the IE with the filtering methods, we used the setup shown in Fig. 4 to collect a tactile dataset that is composed of the response of a single sensor to the three-touch pattern, mentioned before. Fourteen trials have been applied, 5 for TT, 4 for PRT touch and 5 for CT pattern. During each trial, the IE acquires sensor data, applies filtering, and then sends the data to the PC. A LabVIEW GUI collects and stores the tactile data for further analysis.

Filters Coefficients

In the experiments implemented so far, a high noise level was detected in the response of the sensors. This was more observed when plotting the time-Frequency analysis (Continuous Wavelet Transform) of the CT pattern obtained in Fig. 6 and the corresponding charge response (see Fig. 7). The plot shows frequency content of 60 Hz which is the common noise frequency. Moreover, the Time-Frequency plot indicated that the frequency range of the sensor response is below 30 Hz. Such frequency range could be extracted by applying a simple low-pass digital filter with a cutoff frequency of 30 Hz.

The filter coefficients have been computed using MATLAB and saved into the IE local memory. The FIR design was based on Hamming window to create a low pass filter type with a cut-off frequency of 30 Hz. The filter order was set to be 58, the minimum value by which the desired frequencies were filtered out with low hardware computations as possible. As for the EMAF, the α coefficient has been computed by measuring the angular cut-off frequency w_c (Eq. 3) at the half power point of the filter frequency response. Therefore, the corresponding α value for setting 30 Hz as cut-off frequency, was founded to be 0.0909.

$$w_c = \arccos \frac{\alpha^2 + 2\alpha - 2}{2\alpha - 2} \tag{3}$$

3 Performance Analysis

This section provides an experimental evaluation of the sensing system performance. The metrics of interest for our performance analysis are touch detection accuracy, number of reconstructed samples over all trials and filtering execution time.

3.1 Touch Accuracy and Reconstructed Samples

Table 1 reports a comparison between FIR and EMAF in terms of accuracy, number of reconstructed samples and signal-to-noise. The accuracy was computed as the error in detecting the applied touches i.e., difference between the number of actual touches and the one detected by the sensing system. The actual touches have been defined based on the force stimulus detected by the load cell as in Fig. 2 and Fig. 3. Each of the three touch patterns has its corresponding number of applied actual touches i.e., 3 touches for PRT, 10 touches for each of TT and CT. In each trail, a Detection Threshold (DT) was set by enabling the IE to record signals from the sensor for at least 3 secs with no mechanical interaction. Increasing the no touch duration provides more accurate DT. During these three seconds the DT was set by finding the highest (δmax) and the lowest amplitude (δmin) to detect press and release events, respectively (Fig. 8). Both press and release events are considered as detected touches. Table 1 shows that an accuracy of 93.89% was achieved after applying filtering with either FIR or EMAF technique. Furthermore, it shows the impact of filtering on the threshold levels which allows reconstructing a greater number of samples below and above the new thresholds. Results show 12x and 14x increase in the number of reconstructed samples with EMAF and FIR, respectively.

Table 1. Average values of touch detection accuracy, number of reconstructed samples and Signal-to-Noise Ratio (SNR) for all 14 trials collected from the IE

Processing method	No filter	EMAF	FIR
Touch detection accuracy (%)	49.61	93.89	93.89
Number of reconstructed samples	709	8322	10130
SNR (dB)	−5.007	−2.239	−1.944

3.2 Filter Execution Time

Referring to Table 1, FIR filter proved better power and sample reconstruction performance among the EMAF. This would lead, as a conclusion, to selecting FIR filter for implementation on the IE. However, the complexity of FIR is more considerable compared to EMAF, and this is obviously presented in Table 2 by measuring the execution time (ET) and number of clock cycles (CC) preformed for executing both EMAF and FIR filters on the IE. The μVision Debugger supported by ARM Keil framework has been used for debugging, loading, and executing the application program onto the IE.

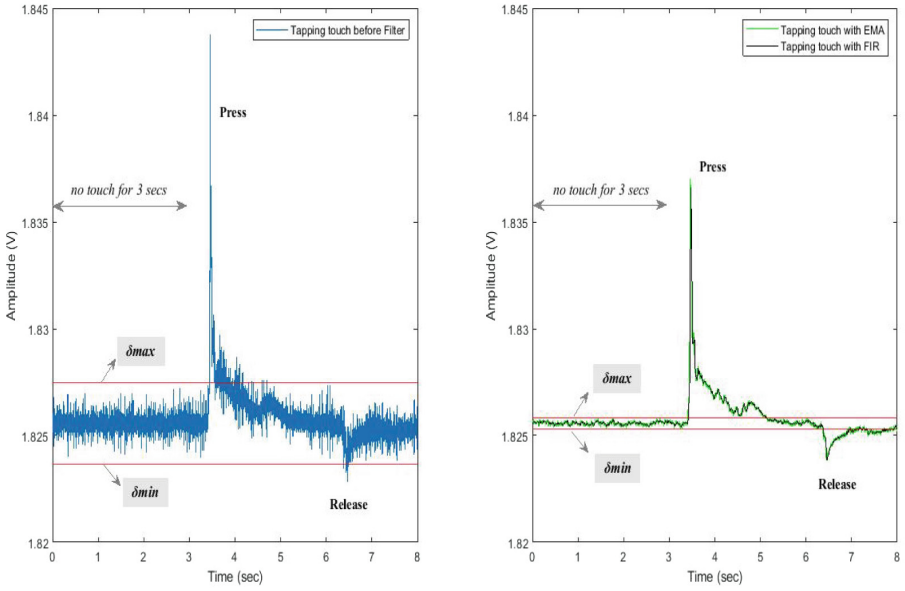


Fig. 8. Max and Min thresholds are set for all collected signals in all trials to detect press and release events; signal before filtering with thresholds (left); signal after filtering (right) one with FIR (black) and one with EMAF (green)

When program execution stops, μVision shows the current state of the CPU registers. Two registers have been monitored to observe the approximated values of ET and CC. As shown in Table 2 the EMAF is 190x faster than the FIR, requiring just 47 clock cycles to be executed. These values correspond to single input sensor; however, the system carries at least 32 input sensors that are sampled simultaneously every 500us. So, with EMAF just 96 us (32×3 us) takes for filtering all the 32 sensors, whereas the FIR execution time is 570 us exceeding the sampling time (500 us) in the IE. Thus, resulting in a loss in some tactile data which consequently reduces the sampling frequency of the system (2 KSample/second).

Table 2. Filtering execution time and clock cycles performed to filter one input signal running onto the IE

Processing method	EMAF	FIR
Filtering execution time (us)	3	570
Clock cycles	47	9120

Therefore, using EMAF will lead to a faster system in acquiring, processing, and transmitting data. In addition to the significant improvements in the quantity of touch samples, which is useful for identifying several levels of touch force.

4 Conclusion

This work presents the implementation of a lightweight pre-processing filtering methods for tactile sensing system. Two filtering methods were considered, FIR and EMAF filters. Both filters were implemented into the IE acquisition board of a complete tactile sensing system. Experimental results showed the effect of filtering methods on the system performance and provides a light-weight pre-processing method based on the EMAF filter compared to the FIR. The sensing system, with EMAF, achieves high touch detection accuracy compared to system running without filtering. In addition to the significant improvement in the number of reconstructed tactile samples, paving the way toward identifying high resolution touch force levels. The future work involves the implementation of more advanced signal processing methods (i.e., machine learning) with the aim of extracting features locally on hardware.

References

1. Johansson, R.S., Flanagan, J.R.: Coding and use of tactile signals from the fingertips in object manipulation tasks. Nat. Rev. Neurosci. **10**(5), 345–359 (2009). https://doi.org/10.1038/nrn2621
2. Saunders, I., Vijayakumar, S.: The role of feed-forward and feedback processes for closed-loop prosthesis control. J. Neuroeng. Rehabil. **8**(1), 60 (2011). https://doi.org/10.1186/1743-0003-8-60
3. Kroemer, O., Lampert, C.H., Peters, J.: Learning dynamic tactile sensing with robust vision-based training. IEEE Trans. Robot. **27**(3), 545–557 (2011). https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5752870&casa_token=T6-SrVedqKgAAAAA:SyWga8VVMbeHb2jlg9btAisVjb6yYbK0_M29RGGB0Zu5gVSMDTqPzvYauEpFCczgdzamv6Wb&tag=1
4. Karim, S.A.A., Kamarudin, M.H., Karim, B.A., Hasan, M.K., Sulaiman, J.: Wavelet transform and Fast Fourier transform for signal compression: a comparative study. In: International Conference on Electronic Devices, Systems and Applications, pp. 280–285 (2011). https://doi.org/10.1109/ICEDSA.2011.5959031
5. Masteller, A., Sankar, S., Kim, H.B., Ding, K., Liu, X., All, A.H.: Recent developments in prosthesis sensors, texture recognition, and sensory stimulation for upper limb prostheses. Ann. Biomed. Eng. **49**(1), 57–74 (2020). https://doi.org/10.1007/s10439-020-02678-8
6. Gianoglio, C., Ragusa, E., Gastaldo, P., Valle, M.: A novel learning strategy for the trade-off between accuracy and computational cost: a touch modalities classification case study. IEEE Sens. J. **22**(1), 659–670 (2022). https://doi.org/10.1109/JSEN.2021.3129323
7. Fares, H., Abbass, Y., Valle, M., Seminara, L.: Validation of screen-printed electronic skin based on Piezoelectric polymer sensors. Sensors **20**(4), 1160 (2020). https://doi.org/10.3390/s20041160
8. Abbass, Y., Saleh, M., Dosen, S., Valle, M.: Embedded electrotactile feedback system for hand prostheses using matrix electrode and electronic skin. IEEE Trans. Biomed. Circuits Syst. **15**(5), 912–925 (2021). https://doi.org/10.1109/TBCAS.2021.3107723
9. Mode, S., Low, B. E. Ble, P. Bl-sa: "BL600 - Datasheet"
10. DDC232, 32-Channel, Current-Input Analog-to-Digital Converter (2004). www.ti.com. Accessed 2 May 2019
11. Saleh, M., Abbass, Y., Ibrahim, A., Valle, M.: Experimental assessment of the interface electronic system for PVDF-based Piezoelectric tactile sensors. Sensors **19**(20), 4437 (2019). https://doi.org/10.3390/s19204437
12. Smith, S.W.: Digital Signal Processing, Chapter 15, pp. 277–284. California Technical Publishing (1999)

Author Index

A
Abbass, Yahya, 727
Aguiar, Paulo R., 379
Al Haj Ali, Haydar, 105
Albanese, Andrea, 137
Albini, Alessandro, 437
Altun, Osman, 205
Alvela Nieto, Maria Teresa, 574
Amin, Youssef, 148
Andersen, Ann-Louise, 597, 616, 626, 667
Andersen, Rasmus, 626
Anguita, Davide, 74
Ansari, Shahid, 276
Apicella, Tommaso, 137
Arzel, Matthieu, 553
Aschenbrenner, Doris, 470
Auer, Sören, 415

B
Baghdadi, Amer, 583
Balletto, Marco, 43
Baruzzi, Valentina, 84
Baumgärtner, Jan, 252
Beckerle, Philipp, 507
Bellotti, Francesco, 564
Benatti, Simone, 160, 518
Benini, Luca, 160
Bergamaschi, Luca, 331
Berta, Riccardo, 564
Bertram, Lukas, 182
Bertrand, David, 553
Beuss, Florian, 405
Bhattacharya, Bishakh, 276
Birgy, Katrin, 53
Blevec, Hugo Le, 543
Bo, Giancarlo, 318
Bode, Dennis, 3
Borinato, Marco, 74
Bosse, Stefan, 362, 387
Bossuyt, Frederick, 607
Bozzi, Alessandro, 425
Bracco, Fabrizio, 636
Briatore, Federico, 225, 235, 310
Brieskorn, Leander, 194
Brink, Michael, 182
Brunelli, Davide, 137
Brunoe, Thomas Ditlev, 597, 626, 667
Bruzzone, Alessandro Arturo, 636, 646
Bulletti, Andrea, 331
Buselli, Irene, 126
Busse, Matthias, 350

C
Cammin, Christoph, 458
Canale, Roberto, 437
Canese, Lorenzo, 65
Cannata, Giorgio, 437
Cansev, Mehmet Ege, 507
Capineri, Lorenzo, 331
Caraveo, Alan, 299
Cardarilli, Gian Carlo, 65
Cardinali, Fabrizio, 318
Carpanzano, Emanuele, 636
Cieslak, Alexandra, 480
Conti, Francesco, 518
Croy, Alexander, 700
Cuniberti, Gianaurelio, 700
Currò, Fabio, 235, 310

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Valle et al. (Eds.): SYSINT 2022, LNNS 546, pp. 737–740, 2023.
https://doi.org/10.1007/978-3-031-16281-7

D
D'Addona, Doriana Marilena, 657
Dabbous, Ali, 215
Dambra, Carlo, 74, 126
Damiani, Cosimo, 331
De Gloria, Alessandro, 564
De Marchi, Luca, 340
Dellepiane, Silvana, 43
Denkena, Berend, 23
Di Nunzio, Luca, 65
Dierend, Hauke, 205
Diguet, Jean-Philippe, 583
Doebbert, Thomas, 458
Donati, Elisa, 518
Donati, Giacomo, 340
Dosen, Strahinja, 530
Doyle, Joseph, 564
Dwivedi, Anany, 507

E
Echelmeyer, Wolfgang, 489
Elogail, Yasmine, 708
Elshazly, Alaa, 708
Estatico, Claudio, 173

F
Fatahi Valilai, Omid, 262, 299
Fazzolari, Rocco, 65
Fedeli, Alessandro, 173
Ferguson, Nikolas, 507
Ferrari, Alessandra, 646
Ferrari, Enrico, 33
Ferraro, Federica, 43
Fitzek, Frank H. P., 470
Fleischer, Jürgen, 252
Fluegge, Wilko, 405
Fough, Nazila, 679, 689
Frisoli, Antonio, 116
Fung, Wai Keung, 679

G
Gallego, Christian Verdonk, 126
Gamal, Heba Allah, 714
Ganesan, Santhosh, 262
Garavagno, Andrea Mattia, 116
Garrone, Andrea, 74
Gastaldo, Paolo, 95
Gelbhardt, Hannes, 574
Gianoglio, Christian, 105, 148
Giovinazzo, Francesco, 437
Gönnheimer, Philipp, 252
Gotz, Reinaldo, 379
Graffione, Simone, 425
Grella, Francesco, 437
Gripon, Vincent, 543, 553

H
Haggag, Ayman, 714
Hannagan, Thomas, 553
Herrmann, Axel S., 362
Herrmann, Kevin, 415
Huang, Shirong, 700

I
Iaconi, Giulia, 43
Ibarlucea, Bergoi, 700
Ibrahim, Ali, 105, 215

J
Junior, Cristiano Soares, 379
Junior, Pedro O. C., 379
Jure, Fabricio Ariel, 530

K
Kahra, Christoph, 415
Kamath, Sanjay, 194
Karrenbauer, Michael, 470
Khalaf, Muhammad K., 708
Kirchheim, Alice, 480
Kjeldgaard, Stefan, 626
Klemme, Heinrich, 23
Klessig, Henrik, 470
Koepler, Oliver, 415
Koerdt, Michael, 362
Krayani, Ali, 13
Krishnamoorthy, Saikiran, 299

L
Lachmayer, Roland, 205, 415
Lang, Walter, 182
Laudante, Gianluca, 497
Lawas, Jonah Marie, 299
Lazzaro, Alessia, 657
Lehmhus, Dirk, 350
Leonardis, Daniele, 116
Léonardon, Mathieu, 543, 553
Lodi, Matteo, 84
Lozano, Pedro, 470
Ludwig, Stephan, 470
Lutzmayr, Dieter, 449

M
Maack, Björn, 362
Madadnia, Behnam, 607
Marcenaro, Lucio, 13
Marin, Pablo, 13
Martin, David, 13
Martinez, Miguel Garcia, 126
Mastronardi, Valentina, 95
Mazzino, Nadia, 74
Mazzoni, Benedetta, 160
Medini, Khaled, 667

Merenda, Massimo, 657
Mete, Mert, 489
Minisi, Simone, 74
Moglianetti, Mauro, 95
Monson, Paulo M. C., 379
Mosca, Marco, 225, 235, 310
Mosca, Roberto, 225, 235, 310
Moser, Gabriele, 43
Mostafa, Hassan, 708, 714
Mounir, Mario, 708
Mozgova, Iryna, 205, 415
Muselli, Marco, 33

N
Napoleone, Alessia, 626, 646
Nehrke, Lara, 480
Neumann, Simone, 480
Nguyen, Tuan, 489
Nielsen, Kjeld, 597, 667
Nielsen, Morten Skogstad, 597, 667
Nozari, Sheida, 13
Nürnberger, Florian, 415

O
Ohlendorf, Jan-Hendrik, 574
Oliveri, Alberto, 84
Oneto, Luca, 74, 126
Orlandi, Mattia, 518

P
Pagano, Saverio, 225
Panes-Ruiz, Luis Antonio, 700
Papa, Federico, 74
Park, HongSeok, 245
Pastorino, Matteo, 173
Pauritsch, Manfred, 449
Pedone, Deborah, 95
Periaswamy, Parthasarathy, 299
Petrini, Laura, 530
Pille, Christoph, 350
Pinna, Nicolò, 33
Pirozzi, Salvatore, 497
Polle, Christoph, 362
Pompa, Pier Paolo, 95
Prabhu, Radhakrishna, 679, 689

Q
Qi, Bowen, 245

R
Ragusa, Edoardo, 95, 137
Rahn, Thomas, 350
Randazzo, Andrea, 173
Ravera, Alessandro, 84
Re, Marco, 65
Redchuk, Andrés, 287
Regazzoni, Carlo, 13
Revetria, Roberto, 225, 235, 310
Riemenschneider, Leif, 700
Rizk, Mostafa, 583
Rodrigues, Alessandro Roger, 379
Rösiö, Carin, 616
Rueckert, Patrick, 53

S
Sacile, Roberto, 425
Saffih, Faycal, 708
Sakr, Fouad, 564
Saleh, Moustafa, 727
Salimbeni, Sergio, 287
Sanchez, Gustavo, 318
Sanetti, Paolo, 74
Scharle, Marvin, 470
Schenone, Valentina, 173
Schiavone, Davide, 518
Schindler, Thimo Florian, 3
Scholl, Gerd, 458
Schotten, Hans D., 470
Sender, Jan, 405
Shah, Chirag, 387
Sheveleva, Tatyana, 415
Skärin, Filip, 616
Slim, Fatima, 583
Sodachi, Majid, 299
Solarna, David, 43
Solzbacher, Bettina, 458
Spaich, Erika Geraldina, 530
Spanò, Sergio, 65
Stepputat, Marten, 405
Stewart, Craig, 679, 689
Stoppel, Dennis, 23
Storace, Marco, 84

T
Taccioli, Tommaso, 137
Taghouti, Maroua, 470
Tagliavini, Giuseppe, 160
Tessier, Hugo, 553
Thoben, Klaus-Dieter, 3, 182, 574
Tonelli, Flavio, 318
Tracht, Kirsten, 53
Trombini, Marco, 43

V
Valencia, Daniel, 194
Valle, Maurizio, 105, 148, 215, 727
Vanfleteren, Jan, 607

Verda, Damiano, 33
Vignola, Giulia, 74
von Hehl, Axel, 387

W
Wawer, Max Leo, 415
Wicaksono, Hendro, 262
Wollnack, Jörg, 194
Wu, Huanzhuo, 470

Y
Younes, Hamoud, 543, 564

Z
Zanghieri, Marcello, 518
Zero, Enrico, 425
Zinn, Carolin, 387
Zonzini, Federica, 340
Zunino, Rodolfo, 95