Contextual Operating Room Monitoring

What pixels tell us about workflow

Butler, R.M.

DOI
10.4233/uuid:8f3f45c5-0c73-4534-b866-2bb59bae8466

Publication date
2025

Document Version
Final published version

Citation (APA)
Butler, R. M. (2025). *Contextual Operating Room Monitoring: What pixels tell us about workflow*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:8f3f45c5-0c73-4534-b866-2bb59bae8466

Contextual Operating Room Monitoring

What pixels tell us about workflow

Rick M. Butler

Contextual Operating Room Monitoring

What pixels tell us about workflow

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus,
Prof.dr.ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
wednesday 24 September 2025 at 12:30 o'clock

by

Rick Maarten BUTLER

Master of Science in Electrical Engineering,
Eindhoven University of Technology, the Netherlands,
born in Eindhoven, the Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof.dr. J.J. van den Dobbelsteen,	Delft University of Technology, *promotor*
Prof.dr. M. van der Elst,	Delft University of Technology, *promotor*
Prof.dr. B.H.W. Hendriks,	Delft University of Technology, *promotor*
Independent members:	
Prof.dr. J. Dankelman,	Delft University of Technology
Prof.dr.ir. D. Ruijters,	Eindhoven University of Technology
Dr.ir. J. Dijkstra,	Leiden University Medical Center
Dr. R.M.A. van der Boon,	Erasmus University Medical Center
Reserve member:	
Prof.dr.ir. T.J.H. Vlugt,	Delft University of Technology

Keywords: Cardiac Catheterisation, Perioperative Workflow, Computer Vision, Pose Tracking, Medical Process Engineering, Future Operating Room, Context-Awareness

Printed by: Ridderprint

Cover by: T.P. van den Hoofdakker

No AI-generated text, images, tables, equations or other materials were included in this dissertation.

The dissertation was written in the Novanext LaTeX template, available at https://gitlab.com/novanext/tudelft-dissertation.

ISBN 978-94-6384-834-3

An electronic copy of this dissertation is available at https://repository.tudelft.nl/.

CONTENTS

LIST OF ABBREVIATIONS

AlphaP AlphaPose

Cath Lab Cardiac Catheterisation Laboratory

ID Identifier

kne Knee

l Left

lbw Elbow

mov Movement

nkl Ankle

OpenP OpenPose

OpenPP OpenPifPaf

pos Position

Pred Prediction

px Pixel

r Right

ReID Reidentification

shr Shoulder

^{T} Tracking algorithm

wrt Wrist

LIST OF ACRONYMS

2D Two-dimensional

3D Three-dimensional

AA Association Accuracy

AI Artificial Intelligence

AP Average Precision

CAF Composite Association Field

CAG Coronary Angiogram

CIoU Complete-Intersection over Union

CNN Convolutional Neural Network

COCO Common Objects in Context

CV Computer Vision

CVAT Computer Vision Annotation Tool

DA Detection Accuracy

EDE Euclidean Distance Error

FN False Negative

FP False Positive

fps Frames Per Second

GCN Graph Convolutional Network

GIoU Generalised-Intersection over Union

GPU Graphical Processing Unit

GT Ground Truth

HAR Human Action Recognition

HKK Hartkatheterisatiekamer

HOTA Higher-Order Tracking Accuracy

IoU Intersection over Union

KI Kunstmatige Intelligentie

LSTM Long Short-Term Memory

MIS Minimally Invasive Surgery

ML Machine Learning

MLP Multi-Layer Perceptron

MOT Multi-Object Tracking

MPII Max Planck Institute for Informatics

NN Neural Network

NW North-West

OK Operatiekamer

OKS Object Keypoint Similarity

OR Operating Room

OS Open Surgery

PAF Part Affinity vector Field

PCKh Head-guided Percentage of Correct Keypoints

PnP Perspective-n-Point

pp Percentage Point

RAS Robot-Assisted Surgery

RdGG Reinier de Graaf Gasthuis

RPE Reprojection Error

S South

SE South-East

SIoU SCYLLA-Intersection over Union

SLP Single-Layer Perceptron

SW South-West

TP True Positive

ViT Visual Transformer

YOLO You Only Look Once

LIST OF MATHEMATICAL SYMBOLS

$\boldsymbol{a}^{(f)}$ Acceleration vector on frame f

A Activity or workflow phase

$\mathcal{A}$ Set of activities or workflow phases

$\boldsymbol{a}$ Annotation vector

A_w Activity or workflow phase in window w

α Activation function

$\mathcal{B}$ Set of bins

$b_{\rho,n}$ Magnitude bin number n

B_ρ Number of magnitude bins

$\mathcal{B}_\rho$ Set of magnitude bins

$b_{\theta,n}$ Orientation bin number n

B_θ Number of orientation bins

$\mathcal{B}_\theta$ Set of orientation bins

β_a Acceleration memory factor

β_j Jerk memory factor

c Detection confidence

$c^{(f_0)}_{p_0,K_0}$ Detection confidence of $d^{(f_0)}_{p_0,K_0}$

$c^{(f)}_{p,K}$ Keypoint detection confidence of class K, belonging to pose p, on frame f

d (Keypoint) detection

$d^{(f)}_{p,K}$ Keypoint detection of class K, belonging to pose p, on frame f

$d_{p_0,K_0}^{(f_0)}$ Reference with respect to which the displacement of $d_{p,K}^{(f)}$ is measured

$d_0^{(f)}$ Short notation for $d_{p_0,K_0}^{(f_0)}$, where p_0 and K_0 are implied and f_0 is a function of f

$d^{(f)}$ Detection on frame f, or short notation for $d_{p,K}^{(f)}$ where p and K are implied

D Number of detections

$\mathcal{D}$ Set of detections

$d_{\text{mov}}^{(s_{\text{mov}})}$ A keypoint detection for staff movement classification within subpose s_{mov}

$d_{\text{pos}}^{(s_{\text{pos}})}$ A keypoint detection for staff position classification within subpose s_{pos}

δ_x Horizontal distance between predicted- and ground truth bounding box centers

δ_y Vertical distance between predicted- and ground truth bounding box centers

Δf The number of frames over which keypoint displacement is measured

η Learning rate

f Frame

F Window length in frames

f_0 Reference frame for history vector calculation

$\mathcal{F}$ Set of frames

f_{mem} The amount of detection-less frames after which BYTE gives up on a tracklet

F_{min} Minimum tracklet length in frames

F_{prior} Number of frames over which workflow phase priors are calculated

γ_{high} BYTE threshold to separate high- and low-confidence detections

γ_{kp} Keypoint confidence threshold for drawing

$\gamma_{\text{mov}}^{(s_{\text{mov}})}$ Keypoint detection confidence threshold to count towards staff movement classification

$\gamma_{\text{pos}}^{(s_{\text{pos}})}$ Keypoint detection confidence threshold to count towards staff position classification

h_{GT} Ground truth bounding box height

$\boldsymbol{h}$ History vector

$h^{(f)}$ History vector element on frame f

$\mathcal{H}$ Set of history subvectors

h_{Pred} Predicted bounding box height

$\boldsymbol{h}_w$ History vector in window w

$\boldsymbol{i}$ Input vector

$\boldsymbol{j}^{(f)}$ Jerk vector on frame f

k Keypoint

K Keypoint class

K_0 Reference keypoint class for history vector calculation

$\mathcal{K}$ Set of keypoint classes

$k_{p,K}^{(f)}$ Keypoint of class K, belonging to pose p, on frame f

L Layer

$\mathcal{L}$ Loss

λ Objective function weight

$\mathcal{M}_A$ Set of Markov models belonging to phase A

M Markov model

$M_{\text{keypoint}}^{s_{\text{mov}}}$ The number of keypoints that must not move before subpose s_{mov} is classified as still

M_{subpose} The number of subposes that must be still for the entire pose to be classified as still

$\boldsymbol{o}$ Output vector

ω Angle between predicted- and ground truth bounding box centers

$P_{\text{keypoint}}^{(s_{\text{pos}})}$ The number of keypoints that must be by the operating table before subpose s_{pos} is classified as such

P_{subpose} The number of subposes that must be near the operating table for the entire pose to be classified as such

p Pose

p_0 Reference pose for history vector calculation

$\mathcal{P}$ Set of all poses

$\boldsymbol{p}^{(f)}$ Position vector on frame f

$\boldsymbol{\phi}$ Machine learning parameters

Q Permutation matrix

R Affinity matrix

$r_{n,m}$ Affinity score between detections n and m

r History vector reference function

$r_{n,m}^{e}$ Epipolar affinity score between detections n and m

ρ_{first} Upper bound of the first magnitude bin

ρ_{last} Upper bound of the last magnitude bin

ρ Magnitude

$r_{n,m}^{t}$ Tracking affinity score between detections n and m

s_{mov} Subpose for movement constraints

s_{pos} Subpose for positional constraints

S Window start interval in frames

σ_{high} BYTE minimum similarity score for matching high-confidence detections

σ_{low} BYTE minimum similarity score for matching low-confidence detections

σ_{new} BYTE minimum similarity score for confirming new tracklets

σ_t Transition probability standard deviation

T_{min} Minimum number of tracklets

$\tau_{\text{mov}}^{(s_{\text{mov}})}$ Threshold below which a keypoint displacement is classified as still

τ_{OKS} Object keypoint similarity threshold when calculating average precision

θ Orientation

$\boldsymbol{u}$ Weight vector

U Weight matrix

v Bias

$\boldsymbol{v}$ Bias vector

$\boldsymbol{v}^{(f)}$ Velocity vector on frame f

$\mathcal{W}_A$ Batch of windows, annotated as phase A

w_{GT} Ground truth bounding box width

w_{Pred} Predicted bounding box width

w Window

$\boldsymbol{x}^{(f)}$ Kalman filter state on frame f

ζ Weight to balance epipolar- and tracking affinity

SUMMARY

Modern healthcare struggles as the population ages and personnel becomes more scarce. Operating rooms (ORs) need to adhere to strict standards and present a major cost for hospitals. Safety of the patient should always be the main concern, and wellbeing of the staff must not be forgotten. Patients have to deal with emotional hardship, delays, and sometimes rescheduling of their treatment. Healthcare professionals experience high workloads due to personnel shortages, culture and new technologies changing the working environment. Hospital management is confronted with the resulting high turnover rates, whilst serving society with their limited available resources.

Efforts are made to improve this situation by assisting hospital employees in their work. New tools can ease tasks, and make them more efficient or safer. Alternatively, finding and teaching optimal ways of working improves workflow efficiency. The desired outcome is to compensate personnel shortages by decreasing the load on employees, whilst maintaining availability and quality of care.

Supportive healthcare tools are often implemented using technology. For example, artificial intelligence (AI) helps find disease in medical imaging, and may assist procedure planning. Devices enable e.g. clear X-Ray imaging with minimal radiation, and steady robotic surgery. However, technological changes to the perioperative setting can burden healthcare professionals in some situations. Beside the patient, they now need to pay attention to these devices. Malfunctions can delay procedures, cause risks for the patient, and induce stress in the staff.

Education through workflow analysis- and optimisation presents an unintrusive path to efficiency improvements. Where process optimisation is already mature in industry, these techniques do not translate directly to hospitals. Healthcare tasks cannot be divided and simplified as in some industries, and every patient has unique needs. However, knowledge-based feedback and support can still be beneficial. Analysing on scale how specialists work allows to extract best practices for safety, efficiency and wellbeing. For example, workflow insights can reveal optimal procedure scheduling, approaches to a workflow phase, adaptation to different patients, and use of new technologies. A current challenge therefore is to formulate scalable workflow analysis methods in hospitals.

Automation through algorithms is a promising tool for scalable workflow analysis, where recorded data from the OR can be translated to workflow metrics. Available datastreams include (monitoring) videos,

device logs, and diagnostic measurements. Differences in protocol and workflow preference between hospitals and medical teams present difficulties. Perioperative workflow analysis must be robust against such variability, recognising relevant patterns regardless of the team or OR. One technology that excels at such robust pattern recognition is deep learning.

Datastreams from the OR present a tradeoff between scope and generalisability. For example, devices may keep logs, which generalise well between ORs, but have a small scope as only device usage is recorded. Although monitoring videos generalise poorly due to visually unique ORs and viewpoints, their view of the whole room yields a large scope. This dissertation investigates the use of monitoring videos for generalisable workflow analysis. Cameras were mounted on the ceiling in a cardiac catheterisation laboratory (Cath Lab) and several ORs. The Cath Lab is a special OR for minimally invasive cardiac interventions, which high level of standardisation presents opportunities for explorative workflow study. Videos of several hundreds of real interventional procedures were recorded. Computer vision (CV) algorithms were used to extract visual- and workflow features.

Part I of this dissertation investigates the extraction of visual features from video data. It tests several CV techniques, describing the challenges of each within the complex environment of the Cath Lab. CV is complicated by occlusion and reflections, which are abundant in the Cath Lab and ORs. In addition, the sterile clothes that are worn make individuals difficult to distinguish. Camera systems bottleneck performance, as synchronisation, sensor noise, and lighting adaptivity leave room for improvement. Generalisability, although desired for the deployment of systems across rooms and hospitals, proves a major challenge. Not many training data are available, viewpoints vary between rooms, and hospitals use different versions of the same object. Viewpoint-dependence can be mitigated by integrating information from multiple calibrated cameras. Although work exists on generalisable feature extraction, in this dissertation it was decided to use human poses, which generalise relatively well.

Part II continues with workflow feature extraction from detected human poses. It is investigated how specific movements of individuals or teams indicate workflow events. Although the used method yielded no usable workflow event detector, results suggested that movements of different bodyparts indeed signal different aspects of workflow. Another approach classified personnel actions by thresholding their position and speed. Workflow differences were measured between procedures with varying involvement of technology. Although both these investigations are exploratory, they demonstrate the potential of context-aware systems which may understand- and collaborate with the staff.

SAMENVATTING

De gezondheidszorg kampt met een vergrijzende populatie en personeelsschaarste. Operatiekamers (OKs) moeten voldoen aan strenge eisen en zijn een grote kostenpost. De veiligheid van de patiënt is altijd de hoogste prioriteit, en het welzijn van werknemers mag niet worden overzien. Patiënten krijgen te maken met confronterende emoties, vertragingen, en soms het verschuiven van hun behandeling. Zorgverleners ervaren hoge werkdruk door personeelstekorten, werkcultuur en het veranderen van hun werkomgeving door nieuwe techniek. Het ziekenhuisbestuur krijgt te maken met de resulterende hoge circulatie van personeel, terwijl ze de maatschappij dient met beperkte middelen.

Er wordt getracht deze situatie te verbeteren door ziekenhuispersoneel ondersteunende hulpmiddelen te bieden. Deze hulpmiddelen kunnen taken vergemakkelijken, en efficiënter of veiliger maken. Ook het vinden en instrueren van optimale werkwijzen verbetert de efficiëntie. Het einddoel is om personeelstekorten te compenseren door de werkdruk op het personeel te verlagen, terwijl de beschikbaarheid en kwaliteit van de zorg behouden blijven.

Hulpmiddelen in de zorg worden vaak geïmplementeerd met technologie. Zo kan kunstmatige intelligentie (KI) helpen bij het vinden van ziektes in medische beelden of het plannen van behandelingen. Moderne apparaten maken scherpe Röntgen beelden met minimale bestraling, of faciliteren stabiele robotoperaties. Echter maken nieuwe technieken het zorgverleners soms lastig in de OK. Naast de patiënt moeten ze zich nu bezig houden met machines. Storingen kunnen interventies verlengen, risico's veroorzaken voor de patiënt, en leiden tot stress bij personeel.

Het analyseren van werkwijze voor opleidingsdoeleinden belooft verbeterde efficiëntie, zonder de behandelomgeving te veranderen. Waar procesoptimalisatie al volwassen is binnen de industrie, is dit niet direct toepasbaar in ziekenhuizen. Zorgtaken zijn niet makkelijk te verdelen en versimpelen zoals in sommige industriën, en elke patiënt heeft diens eigen behoeften. Door inzicht gedreven feedback en ondersteuning zijn echter nog steeds wenselijk. Door op schaal te analyseren hoe experts te werk gaan kunnen optimale werkwijzen gevonden worden voor veiligheid, efficiëntie en welzijn. Zulke inzichten onthullen bijvoorbeeld optimale plannings, de beste aanpak van een zorgtaak, aanpassingen per type patiënt, of correct gebruik van nieuwe techniek. Een uitdaging is nu hierom het schaalbaar analyseren van werkwijze in ziekenhuizen.

Algorithmes kunnen data uit OKs schaalbaar analyseren, en vertalen

naar metrieken voor werkwijze. Mogelijke datastromen omvatten (surveillance) video's, logboeken, en medische beelden. Verschillende protocollen, en voorkeuren van ziekenhuizen en behandelaars, bemoeilijken analyses. Een algoritme moet onder deze variabiliteit nog steeds patronen kunnen herkennen, ongeacht de OK of behandelaar. Een veelbelovende techniek hiervoor is deep learning.

Datastromen uit de OK bieden een afweging tussen bereik en generalisering. Sommige apparaten houden bijvoorbeeld logboeken bij, welke goed generaliseren naar andere OKs met hetzelfde apparaat, maar beperkte informatie bevatten. Surveillance video's generaliseren slecht doordat elke OK en kijkhoek uniek is, maar ze zien wel alles dat gebeurt. In deze dissertatie wordt het gebruik van surveillance voor het analyseren van werkwijze onderzocht. Hiervoor zijn camera's opgehangen in een hartkatheterisatiekamer (HKK) en meerdere OKs. De HKK is een speciale OK, welke is ingericht voor minimaal invasieve hartdiagnostiek. De hoge standaardisatie van deze ingrepen maakt de HKK erg geschikt voor een verkennende studie naar werkwijze. Enkele honderden echte interventies zijn gefilmd. Computervisie (CV) algoritmes analyseerden vervolgens deze beelden om informatie over werkwijze te verkrijgen.

Deel I van deze dissertatie onderzoekt het automatisch herkennen van visuele patronen in video's uit de HKK. Er worden meerdere CV methodes getest, en uitdagingen in de HKK worden omschreven. Occlusie en reflecties, welke veel voorkomen in de HKK en OKs, bemoeilijken CV. Steriele kleding maakt het moeilijk om personen te onderscheiden. Camerasystemen begrenzen de prestaties door gebrekkige synchronisatie, aanpassingsvermogen aan verlichting, en sensor ruis. Generaliseerbare systemen zouden helpen bij toepassingen op schaal, maar dit blijkt een grote uitdaging. Beelden uit het ziekenhuis zijn slecht beschikbaar, cameraperspectieven verschillen per HKK en OK, en ziekenhuizen gebruiken verschillende versies van dezelfde objecten. De afhankelijkheid van cameraperspectief wordt opgelost door informatie uit meerdere camera's te combineren. Ondanks bestaand onderzoek over generalisatie in CV wordt in deze dissertatie besloten om enkel mensen te detecteren, welke er vergelijkbaar uitzien in verschillende situaties.

In Deel II wordt werkwijze geanalyseerd aan de hand van gedetecteerde mensen. Er wordt onderzocht hoe specifieke bewegingen van personen of teams informatie over werkwijze bevatten. De methode bleek niet bruikbaar voor het classificeren van werkwijze, maar impliceerde wel dat verschillende lichaamsdelen informatie bevatten over verschillende aspecten van werkwijze. Een ander onderzoek classificeerde de acties van personeel aan de hand van hun positie en bewegingssnelheid. Deze methode mat verschillen in werkwijze tussen procedures van verschillende technische aard. Hoewel deze onderzoeken verkennend waren, demonstreren ze de mogelijkheden van een context-bewust systeem welke de acties van personeel begrijpt en met ze samenwerkt.

PREFACE

After obtaining my Master's degree in electrical engineering, I felt like there was a great deal left to learn. This strong desire to keep learning ultimately lead to my application for Ph.D. candidacy. My own experiences as a patient, and an ambition to contribute to societal wellbeing, led me to the field of healthcare technology.

As an engineer among clinicians, I learned not only about healthcare technology, but also of the importance of effective communication. Trust and collaboration between engineers and clinicians is the basis of lifting healthcare to new heights. Misunderstandings of each other's needs and abilities form a gap between these two groups. With this thesis, I hope to illuminate some challenges in healthcare for engineers, and in technology for clinicians. In today's struggling healthcare society, bringing these two groups together is more important than ever.

Rick Maarten Butler
Delft, April 2025


1

INTRODUCTION

This chapter serves as background for the rest of the book. A brief overview is presented on the history of process optimisation, cardiac catheterisation and machine learning. For each of these topics, their origin, development, and current status are discussed. The inner workings of machine learning and deep learning, and their application to computer vision, are briefly addressed. The chapter concludes with the research question and its subquestions, that the rest of this book aims to answer.

1

Personnel shortages and growing demands in healthcare have been worldwide challenges for years [1]. This situation leads to limited or unequal availability of care, and a rising workload among healthcare professionals [2]. In the year 2020, the global health workforce shortage was estimated at 15.4 million workers [1]. Developed countries source healthcare professionals from developing countries as a temporary remedy. This contributes to 47 % of global healthcare being reserved to 22 % of the world population in 2020. Although the healthcare workforce is expected to grow in the short term, one third of current medical doctors are expected to retire within fifteen years [3]. Healthcare shortages and the resulting delays cause emotional distress and anxiety in patients [4]. Resulting patient hostility and high workloads cause feelings of anger, frustration, anxiety and worry in healthcare workers [5]. This has been known to negatively affect care quality, delays, job satisfaction, and to induce risk avoidance through e.g. redundant treatment and overprescription.

Proposed solutions to healthcare workload shortages include the education and recruitment of additional staff, or the development of tools to assist them in their work. The latter approach has been pursued through new technologies and process optimisation [6]. Effective assistance with medical tasks reduces staff workload, and improves the safety and efficiency of patient care. However, hospitals and operating rooms are a complex environment. New technologies may have unforeseen effects on the staff or patient, which should be investigated before their implementation.

Recently, artificial intelligence (AI) has received much attention as a healthcare support tool [7]. The history and implementations of AI are further discussed in section 1.3. AI has assisted with diagnosis, medical imaging, hospital management, workflow analysis, and even communication with patients. However, it is important to be aware of its limitations. AI reasoning is fundamentally different from that of a human being. It does not have the ability to adapt to new, unexpected situations that it never encountered during training. Therefore, in its current state, AI cannot be relied upon as an independent agent. Rather, its role should be to assist medical staff, providing insights and suggestions. The responsibility for patient care must remain in the hands of healthcare professionals.

This dissertation concerns the use of AI for workflow analysis in the cardiac catheterisation laboratory (Cath Lab) and other operating rooms. Specifically, monitoring footage is analysed to obtain workflow information. The background provided in this first chapter lays the foundation for the rest of the book. Section 1.1 starts with a brief history of workflow optimisation in industry, and its extension to healthcare. Industry and healthcare pursue different goals, and their consequently different workflow requirements are emphasised. Workflow is relatively

well-defined during cardiac catheterisation in the Cath Lab, which is described in section 1.2. Before considering the application of AI to workflow analysis in the Cath Lab and other operating rooms, it is important to understand its basic inner workings. To this end, section 1.3 provides a brief summary of the fundamentals of AI, and its applications in healthcare. Deep learning is addressed in particular, as this is the state-of-the-art approach to AI at the time of writing. AI for surgical monitoring and workflow analysis are highlighted, as this application is the major focus of this dissertation. Section 1.4 concludes this chapter by stating the research question and subquestions, that the rest of this work aims to answer.

1.1. PROCESS OPTIMISATION

"The great doesn't happen through impulse alone, and is a succession of little things that are brought together." [8]

Running a steel production workshop around 1878, F. W. Taylor hypothesised that the work could be streamlined through study and organisation [9]. His 'scientific management' divided complex tasks into simpler jobs, separated planning from execution, and it centralised knowledge and decision-making [10]. H. L. Gantt—his assistant—promoted positive incentive, employee responsibility, and job satisfaction [11, 12]. In the 1910s, H. Ford proposed the assembly line to standardise tasks and work pace in his automotive company [13]. Several process optimisation methods evolved in industry afterwards [14]. Two popular methods are Lean and Six Sigma, which reduce process waste and variability since the 1980s.

Taylorism was brought into healthcare in the 1930s by dividing responsibilities within the nurse profession [15]. By 1952 nursing practice was standardised across the United States. In the 60s and 70s, hospitals started standardising procedures (inter)nationally to ensure similar outcomes in every hospital [16]. Lean Six Sigma was first implemented in healthcare in a Dutch hospital in 2005 [17]. A variety of studies followed, demonstrating increased patient capacity and satisfaction, and reduced prescription errors, administrative tasks, and waiting times [18]. The COVID-19 pandemic from 2019 to 2023 increased hospital workloads drastically, renewing interest in healthcare process optimisation and Lean Six Sigma [18–20]. Research shifts more and more towards task automation and support systems [19]. Simulation is used to analyse and optimise processes—possibly using AI—before applying changes in practice [21, 22]. Results yield improvements in scheduling, perioperative workflow, OR performance metrics, and hospital flows.

Advocates and critics of process optimisation in healthcare debate the added value and effects on personal care, flexibility and job satisfaction [23, 24]. Indeed, healthcare is very different from industry. The main challenge therefore is to improve healthcare processes in a way that does not negatively affect the quality and availability of care.

1.1.1. Interventional workflow

Hospitals are very different from industry, where process optimisation initially developed. Unlike some industrial production processes, patient care in operating rooms is a nonlinear and complex process that revolves around human beings. Patient safety and comfort must remain the top priority at all times. Hospitals and staff initially demonstrated hesitation towards interventional workflow optimisation, as attempted industrialisation of healthcare was a feared side effect.

Over time, rising demands from an aging society and personnel shortages motivated acceptance and appreciation of medical process engineering. Workflow optimisations in hospitals and healthcare can be introduced under the strict constraints that patient safety and comfort do not suffer. Hospitals started adopting workflow optimisation using e.g. Lean Six Sigma to increase care efficiency and safety [6, 19]. Process optimisations were studied in procedure planning and during procedures in the operation room, which has a large financial footprint compared to other hospital departments.

Medical interventions, staff, and patients can be highly variable, with unpredictable emergencies and complications. Measuring and understanding interventional workflow at scale is an active field of research [25, 26]. Sensors include radio-frequency trackers, kinematic sensors, and (monitoring or in vivo) cameras. Algorithms extract workflow information from measurements. AI has been applied to help with e.g. screening and diagnosis, steadiness of robotic surgery, and presenting important sensor information to healthcare professionals during surgery [27]. However, truly automated perioperative contextual understanding is still far away.

The high variability and many edge cases within surgeries complicate the development of experimental workflow analysis systems. For such pilot studies, it is useful to start in a stable environment with little deviation from protocol. A well-defined procedure with little risk of complications is the coronary angiogram (CAG) [28]. During a CAG, heart problems are revealed by administering a contrast fluid and making X-ray recordings. This diagnostic cardiac intervention is carried out in a specialised operating room: the cardiac catheterisation laboratory (Cath Lab).

1.2. THE CARDIAC CATHETERISATION LABORATORY

"The catheterisation laboratory today combines [cardiac] diagnosis and therapeutics, through various imaging modalities and a prolific list of interventional tools." [29]

In 1929, W. Forssmann—a small-town general practitioner—got obsessed with the idea of administering medicine directly into the heart [30]. He guided a urological catheter to the right atrium of his own heart through the brachial vein, which starts near the inner elbow [31]. A brief overview of the heart atonomy is presented in fig. 1.1, and explained more thoroughly in [32, Chapter 16.1].

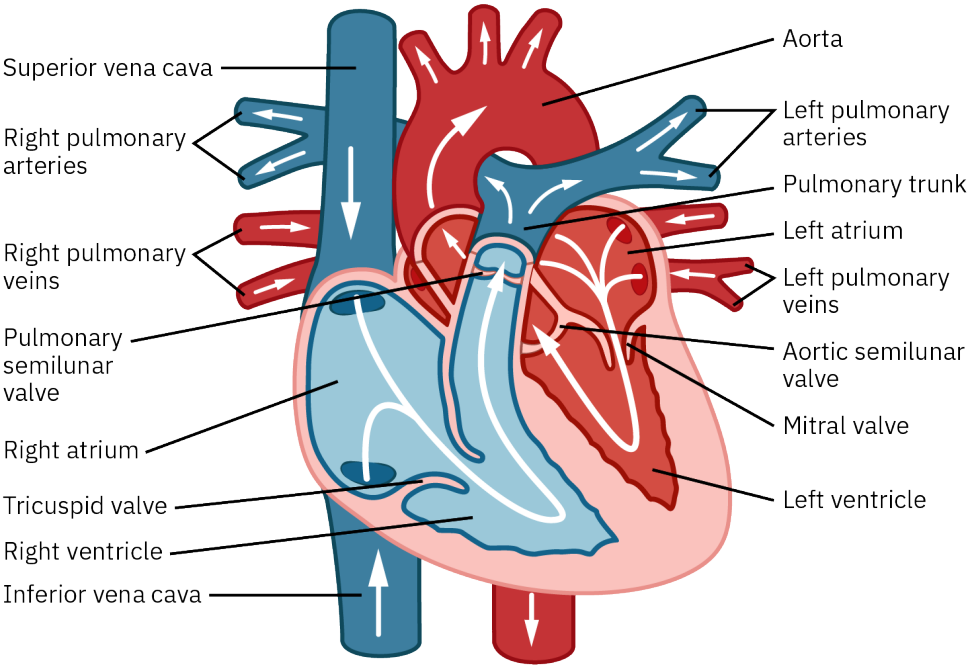


Figure 1.1.: A schematic of the heart [32, Chapter 16.1]. Deoxygenated blood arrives through the venae cavae into the right atrium. The blood is pumped through the right ventrical and the pulmonary arteries to the lungs. Freshly oxygenated blood arrives through the pulmonary veins into the left atrium, and is pumped through the left ventricle into the aorta. The aorta branches off to transport oxygenated blood throughout the body. Coronary arteries (not shown) also branch from the aorta to supply the heart muscle with oxygenated blood. (attribution: [32] © Rice University, OpenStax under CC BY 4.0 license)

Inspired by Forssmann, the first ever Cath Lab was opened in 1945

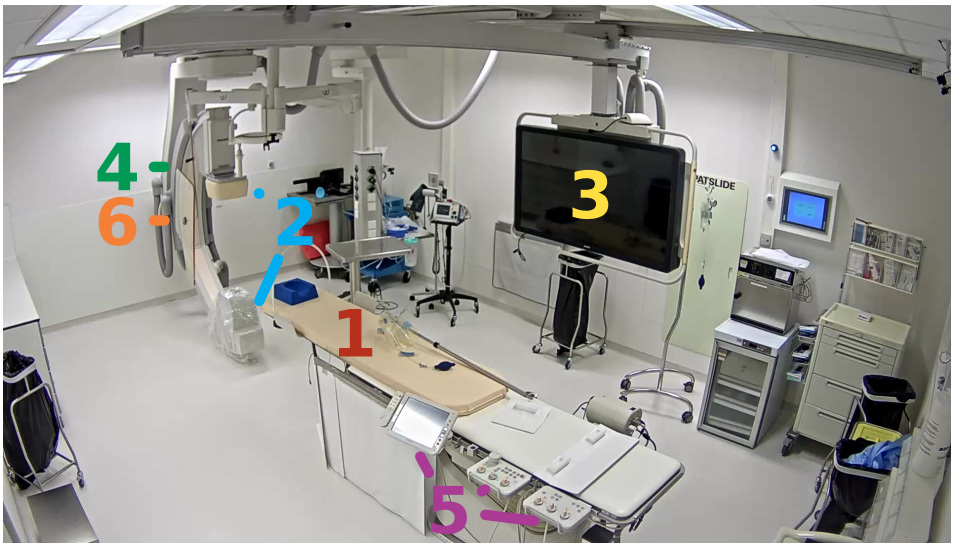


Figure 1.2.: The Cath Lab of the Reinier de Graaf Gasthuis hospital, Delft, NL. Cath Lab equipment is marked: 1) the operating table, 2) the X-ray device, 3) the monitor, 4) the C-arm, 5) the control panel, and 6) the lead shield.

by A. Cournand and D. W. Richards [33]. They designed catheters specifically for cardiovascular interventions [34]. Right—and later left—heart catheterisation became well-documented. Repeated attempts to catheterise the coronary arteries—which supply the heart muscle itself with oxygened blood—remained risky [35]. In 1958, F. M. Sones accidentally administered contrast fluid into the right coronary artery during an aortic catheterisation. Sones' method was born [36], and still forms the basis of CAGs today.

Sones continued to research optimal CAG protocols, and collaborated with Philips North America to develop CAG-specific imaging equipment [35]. Innovations to robotically move the patient and X-ray device enabled convenient imaging from different angles. The introduction of X-ray video recordings enabled computer-assisted analysis and decision making in the 1980s: quantitative CAG [37].

Innovations in the 2000s yielded new imaging methods and improved image quality [29, 38]. The 2010s saw an shift of attention towards radiation-safety [39] and data processing [29, 37, 38]. Personnel radiation exposure was reduced through e.g. shielding, X-ray beamforming, image enhancement, education, and robotic catheterisation. AI developments in the 2020s enabled e.g. multimodal imaging, patient-specific 3D modelling of the heart, and blood flow

simulation for risk prevention. Real-time image analysis and (virtual reality) overlays are currently under investigation to provide guidance during catheterisation.

A modern Cath Lab is shown in fig. 1.2. During a procedure, the patient is placed on the mobile operating table. The X-ray system records images, which are displayed on the monitor. The X-ray source and detector are mounted on a C-arm, which can move and rotate to enable imaging from different angles. The operating table, X-ray source, and C-arm are all controlled by the cardiologist or the technician using a control panel. A mobile lead shield protects staff from radiation. In addition, protective clothing is often worn such as shielding aprons, collars and glasses.

During an intervention, the cardiologist, a scrub nurse, and several assistants are in the room. The assistants bring and prepare the patient before the procedure starts. The staff prepares the necessary sedative, catheter, sheets and instruments. The patient is normally awake, but slightly sedated during the procedure. After sedative is administered, the cardiologist carries out the procedure, controlling the table, C-arm, and X-ray. The scrub nurse stays in the room to provide support, whilst remaining staff observes from the radiation-safe control room. The cardiologist and scrub nurse communicate with the patient to ensure their wellbeing during the procedure. The vitals of the patient are closely monitored in real time on the monitor and within the control room. After the procedure, the assistants re-enter to escort the patient to the recovery area.

1.2.1. THE CORONARY ANGIOGRAM

The purpose of the CAG is to identify obstruction or narrowing of the coronary arteries, which supply the heart muscle with oxygen and other nutrients. First, a sedative is administered to comfort the patient. A catheter is inserted by direct puncture into the radial artery of the wrist, and guided through the aorta into a coronary artery. If the radial artery is too narrow or curved to guide the catheter, access is realised through a puncture of the femoral artery in the groin. Contrast fluid is administered through the catheter directly into the vascular structures of the heart. This fluid reveals the condition of cardiac blood vessels in X-ray images. Thus, vessel obstructions by atherosclerosis or blood clots can be observed. This process is repeated for the left- and right coronary arteries. Finally, the catheter is retracted and the entry wound is closed.

As discussed, the last ten years have shown increasing applications of AI in the Cath Lab, during CAGs and other procedures. Its value becomes especially apparent in fast and high-quality image processing, enabling sophisticated real-time support and assisted diagnostics. The

following section explains the basics of AI, and ends with some recent applications in healthcare.

1.3. ARTIFICIAL INTELLIGENCE

> *"Artificial intelligence (AI) is technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy."* [40]

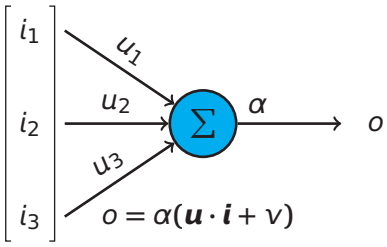
There are many possible ways to implement AI. For example in classical programming, a human provides a computer with a strict set of instructions. Given an input, the computer will execute this program exactly and produce the result. Not all problems are easily solved this way because not all problems have a known, efficient algorithmic description. For example, consider the case where an image needs to be described using a single word. How does one formulate a solution, that holds under a vast variation of object appearances, lighting conditions, camera properties, and viewpoints?

Machine learning (ML) reversed the paradigm. A human provides a set of input-output examples, without specifying their algorithmic relation. In the example above, this would be a set of images with corresponding single-word captions. A machine learns an algorithm by itself, from these examples. Such an algorithm often extracts and refines 'features': descriptions of the input data on varying levels of detail. Low-level features are combined into higher-level features in a number of steps. The human designer imposes constraints—an 'architecture'—within which the machine is free to explore solutions. After learning, the machine should be able to generate accurate outputs for inputs it has never seen before.

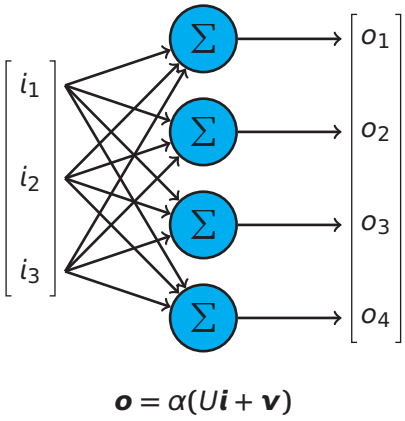
ML is used to solve problems that could not be solved—or could not be solved efficiently—by classical programming. Each ML algorithm solves one specific task, such as recognising images, transcripting speech, designing chemical compounds, or driving cars [41]. ML is made possible by the vast amounts of annotated data generated by modern society [42]. Neither classical- nor ML algorithms are superior to the other; the best approach depends entirely on the problem that one is trying to solve.

1.3.1. DEEP LEARNING

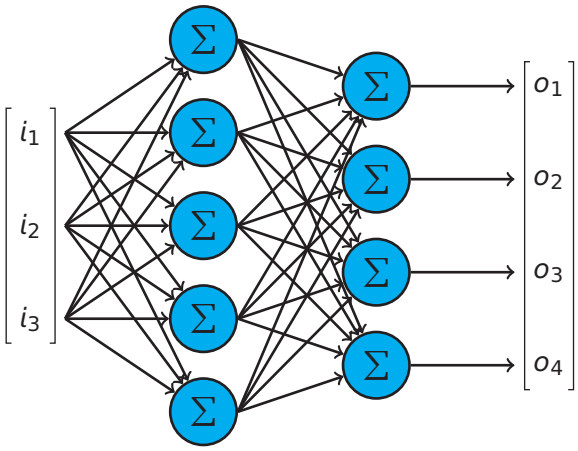
Many ML methods exist, such as support vector machines, logistic regression, and random forests [43]. However, one stands out in terms of learning capabilities: the neural network (NN) [41, 44]. NNs are loosely based on the operation of the brain, where many simple, interconnected 'neurons' form a complex network.



(a) Schematic of a single perceptron neuron. Its transfer function is shown for input vector $\boldsymbol{i}$, weight vector $\boldsymbol{u}$, bias v, and activation function α.

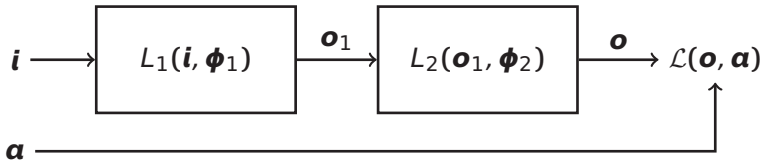


(b) A single-layer perceptron with four neurons. The weights and activation function of fig. 1.3a are present for each neuron, but omitted in this illustration. In the shown transfer function, matrix U and vector $\boldsymbol{v}$ combine the weights and biases of all neurons.

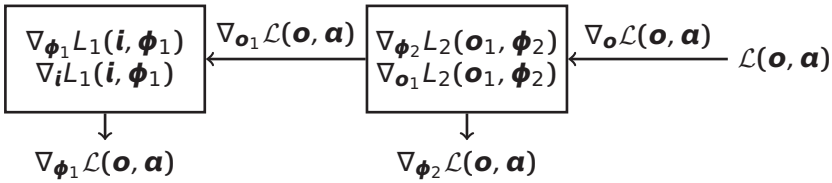


(c) A multi-layer perceptron with one ‘hidden’ layer and one output layer. The first layer has five neurons, and the second has four.

Figure 1.3.: The first steps of neural network development. Perceptrons that share the same inputs collaborate in ‘layers’ to form vector outputs.



(a) The forward pass. Given an input $\boldsymbol{i}$ and corresponding annotation $\boldsymbol{a}$, the network computes an output $\boldsymbol{o}$. The loss function $\mathcal{L}$ computes the dissimilarity between $\boldsymbol{o}$ and $\boldsymbol{a}$. During the forward pass, the output and gradients of each layer L_n with respect to its input and parameters are computed numerically.



(b) The backward pass. The numerical gradient of the loss function with respect to $\boldsymbol{o}$ is 'backpropagated' to obtain the gradient with respect to $\boldsymbol{\phi}$. In each layer n, the chain rule is applied to obtain $\nabla_{\boldsymbol{o}_{n-1}} \mathcal{L}(\boldsymbol{o}, \boldsymbol{a}) = \nabla_{\boldsymbol{o}_{n-1}} L_n(\boldsymbol{o}_{n-1}, \boldsymbol{\phi}_n) \nabla_{\boldsymbol{o}_n} \mathcal{L}(\boldsymbol{o}, \boldsymbol{a})$ and $\nabla_{\boldsymbol{\phi}_n} \mathcal{L}(\boldsymbol{o}, \boldsymbol{a}) = \nabla_{\boldsymbol{\phi}_n} L_n(\boldsymbol{o}_{n-1}, \boldsymbol{\phi}_n) \nabla_{\boldsymbol{o}_n} \mathcal{L}(\boldsymbol{o}, \boldsymbol{a})$. The gradients for all layers combine into $\nabla_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{o}, \boldsymbol{a})$.

Figure 1.4.: Backpropagation in a two-layer NN, to find the numerical gradient of the loss function $\mathcal{L}$ with respect to the network parameters $\boldsymbol{\phi}$. Each layer performs an arbitrary, differentiable vector operation, such as the SLP function of fig. 1.3b. The function applied by layer n is denoted as $L_n(\boldsymbol{x}, \boldsymbol{\phi}_n)$, where $\boldsymbol{x}$ is the layer input and $\boldsymbol{\phi}_n$ the layer parameters. The parameters of all layers together form $\boldsymbol{\phi}$.

The idea of simulating a neuron was first proposed by W. S. McCulloch—a psychologist—and W. Pitts—a logician—in 1943 [45]. In 1958, psychologist F. Rosenblatt generalised their idea to the 'perceptron' [46], which had the ability to learn from data and forms the basis of NNs today. A perceptron—visualised in fig. 1.3a—computes the weighted sum of multiple inputs and adds a bias, before applying a nonlinear activation function [47]. This simulates the firing of a neuron.

Perceptrons produce a single scalar output. By arranging multiple perceptrons as in fig. 1.3b, Rosenblatt realised multiple outputs. This arrangement is called the 'single-layer perceptron' (SLP). As the SLP transfer function is trivial to differentiate, optimal weights and biases could be found through gradient-based optimisation on training data.

Many data contain relations that cannot be modelled using a single linear- and nonlinear vector operation. Rosenblatt published a solution in 1962 [48] with the multi-layer perceptron (MLP), shown in fig. 1.3c. A MLP applies several SLPs consecutively. Here, each SLP is referred to as a network 'layer'. The universal approximation theorem [49] proves that MLPs can approximate any continuous function with arbitrary accuracy, on a finite input domain. However, they often do so in a memory- or computation-inefficient way, justifying exploration of non-SLP layers for specific problems [41, 44]. Referring to multi-layer networks as 'deep', the field of NNs was named deep learning.

The process of training aims to find a set of weights and biases—the parameters—that lets the network approximate a desired input-output relation. This is achieved by minimising a loss function, which measures the difference between produced- and desired network outputs [50]. The backpropagation method—popularised by two psychologists and a computer scientist—enabled gradient-based training of any neural network architecture [51]. As visualised in fig. 1.4, it uses the chain rule to numerically obtain derivatives of a loss function with respect to the network parameters. The parameters are updated by an optimiser algorithm using these derivatives [52]. This numerical optimisation is affected by noise and bias from the training dataset. Additionally, loss functions contain low-gradient regions at non-optimal parameters, halting further optimisation. These effects are mitigated by e.g. averaging gradients over multiple data samples (batch training), large datasets, and carefully designed network architecture, loss functions, and optimisers.

Deep learning was held back by computer hardware limitations and data availability [42]. It took until the 2010s for deep learning implementations to become widespread. In 2012, AlexNet [53]—implemented efficiently on a graphical processing unit (GPU)—presented a major step in image classification technology. Since then, NNs have dominated the field of computer vision, and their development has only accelerated. In healthcare, too, NNs are applied for various purposes.

1.3.2. NEURAL NETWORKS IN HEALTHCARE

Deep learning has numerous applications in healthcare. NNs analyse medical images and records, helping with fast and accurate diagnoses or anomaly detection [7, 41, 43]. They have been applied to recognise signs of disease early on. Generation of synthetic medical images serves educational purposes. Deep learning enables virtual patient care, where patients are monitored using (wearable) technology and receive automated health recommendations. NNs assist in the development of new medication, e.g. by analysing viral proteins or modelling biological processes. They enhance precision during (remote) robotic surgeries [27].

NNs are applied to surgical workflow analysis as well [19, 26]. Here, the main aim is to detect surgical workflow phases and actions at varying granulatities. Laparoscopic videos are the most popular input feature because of their intrinsically rich information [54]. Other inputs such as medical imaging, audio, speech transcriptions, device logbooks, positioning sensors and tool usage have been used as well, albeit less succesfully. Another promising input is monitoring, i.e., surveillance video, recording procedures from a distance [55, 56]. Here, depth- and/or multiview camera systems are popular. Several studies focus on personnel pose detection, where 3D poses are preferred for their generalisability and robustness against occlusion [57, 58]. References [59–61] made the first steps towards surgical workflow analysis from these human pose data. In this dissertation, the optimal ways to analyse surgical monitoring footage—and their applications to workflow analysis—are explored.

1.4. RESEARCH QUESTION

The main question this dissertation aims to answer is:

> *"How can we extract workflow information from Cath Lab monitoring video footage, and use it to improve interventional efficiency?"*

This research question is divided into the following subquestions:

1.4.1. WHICH COMPUTER VISION TECHNOLOGIES CAN BE APPLIED FOR CATH LAB MONITORING?

Automated perioperative video monitoring is a relatively new field. To a computer, a video is nothing more than a collection of pixels without structure or meaning. Therefore, algorithms need to extract meaningful 'features' from video data, creating structure such that computers can understand an image. Chapter 2 provides an overview of popular

computer vision methods, and reviews work on their application in Cath Labs and ORs.

1.4.2. WHICH 2D HUMAN POSE ESTIMATOR PERFORMS BEST WITHIN THE VISUAL COMPLEXITY OF THE CATH LAB?

Cath Labs and ORs present specific challenges such as occlusion, reflections, and varying lighting conditions. Computer vision algorithms designed for general conditions cannot be assumed to perform in the Cath Lab. In workflow analysis, the position and pose of personnel can be a valuable source of information. Chapter 3 benchmarks methods for 2D human pose estimation in the Cath Lab. Cath Lab-specific challenges are highlighted, and an algorithm is recommended that performs best within these constraints.

1.4.3. HOW TO REIDENTIFY PERSONS IN THE CATH LAB?

Detecting persons in still images provides no information on their motion over time. For this, it is necessary to reidentify individuals on different video frames. Challenges from chapter 3—sterile clothing in particular—complicate this computer vision task in the Cath Lab. A solution is explored in chapter 4.

1.4.4. WHICH ASPECTS OF MOTION ARE MOST DESCRIPTIVE OF CATH LAB WORKFLOW?

Personnel motion in the Cath Lab can be an input feature to workflow analysis. Chapter 5 explores workflow phase classification from this feature. Different methods present a tradeoff between performance and interpretability. To analyse the workflow information offered by motion of different body parts, a method is chosen that prioritises interpretability.

1.4.5. HOW CAN PERSONNEL ACTIONS BE CLASSIFIED IN THE OR?

An alternative to the complex motion analysis in chapter 5 is to infer actions simply from personnel positions. Chapter 6 uses this approach for a binary classification of personnel actions in the OR. Here, pose detections are used to compensate for camera perspective. Results are discussed in the light of workflow phase- or event analysis.

Answers to these questions, and their implications, are discussed in chapter 7.

REFERENCES

[1] M. Boniol, T. Kunjumen, T. S. Nair, A. Siyam, J. Campbell, and K. Diallo. “The Global Health Workforce Stock and Distribution in 2020 and 2030: A Threat to Equity and ‘Universal’ Health Coverage?” In: *BMJ Glob. Health* 7.6 (June 2022). doi: 10.1136/bmjgh-2022-009316.

[2] A. Džakula, D. Relić, and P. Michelutti. “Health Workforce Shortage - Doing the Right Things or Doing Things Right?” In: *Croat. Med. J.* 63.2 (Apr. 2022), pp. 107–109. doi: 10.3325/cmj.2022.63.107.

[3] D. T. Michaeli, J. C. Michaeli, S. Albers, and T. Michaeli. “The Healthcare Workforce Shortage of Nurses and Physicians: Practice, Theory, Evidence, and Ways Forward”. In: *Policy, Politics, & Nursing Practice* 25.4 (Nov. 2024), pp. 216–227. doi: 10.1177/15271544241286608.

[4] A. Rodríguez-Fuertes, P. Reinares-Lara, and B. Garcia-Henche. “Incorporation of the Emotional Indicators of the Patient Journey into Healthcare Organization Management”. In: *Health Expect.* 26.1 (Feb. 2023), pp. 297–306. doi: 10.1111/hex.13656.

[5] R. Sattar, R. Lawton, G. Janes, M. Elshehaly, J. Heyhoe, I. Hague, and C. Grindey. “A Systematic Review of Workplace Triggers of Emotions in the Healthcare Environment, the Emotions Experienced, and the Impact on Patient Safety”. In: *BMC Health Serv. Res.* 24 (May 2024), p. 603. doi: 10.1186/s12913-024-11011-1.

[6] A. M. Schouten, S. M. Flipse, K. E. van Nieuwenhuizen, F. W. Jansen, A. C. van der Eijk, and J. J. van den Dobbelsteen. “Operating Room Performance Optimization Metrics: a Systematic Review”. In: *J. Med. Syst.* 47.1 (Dec. 2023), p. 19. doi: 10.1007/s10916-023-01912-9.

[7] A. A. Kuwaiti, K. Nazer, A. Al-Reedy, S. Al-Shehri, A. Al-Muhanna, A. V. Subbarayalu, D. Al Muhanna, and F. A. Al-Muhanna. “A Review of the Role of Artificial Intelligence in Healthcare”. In: *J. Pers. Med.* 13.6 (June 2023), p. 951. doi: 10.3390/jpm13060951.

[8] V. W. van Gogh. “Letter #274 to Theo van Gogh”. In: (Oct. 1882).

[9] P. B. Darmody. “Henry L. Gantt and Frederick Taylor: The Pioneers of Scientific Management”. In: *AACE Int. Trans.* (2007), PS.15.

[10] F. W. Taylor. *Shop Management*. Harper & Brothers, 1911.

[11] J. W. Herrmann. *The Legacy of Taylor, Gantt, and Johnson: How to Improve Production Scheduling*. Tech. rep. The Institute for Systems Research, 2007.

[12] H. L. Gantt. *Organizing for Work*. Harcourt, Brace & Howe, 1919.

[13] D. Watson. "Fordism: A Review Essay". In: *Labor Hist.* 60.2 (2019), pp. 144–159. doi: 10.1080/0023656X.2019.1537031.

[14] A. Chiarini. "Japanese Total Quality Control, TQM, Deming's System of Profound Knowledge, BPR, Lean and Six Sigma: Comparison and Discussion". In: *Int. J. Lean Six Sigma* 2.4 (Nov. 2011), pp. 332–355. doi: 10.1108/20401461111189425.

[15] M. C. Lundy. "Nursing Beyond Fordism". In: *Empl. Responsib. Rights J.* 9.2 (June 1996), pp. 163–171. doi: 10.1007/BF02622257.

[16] M. Berg. "Problems and Promises of the Protocol". In: *Soc. Sci. Med.* 44.8 (Apr. 1997), pp. 1081–1088. doi: 10.1016/S0277-9536(96)00235-3.

[17] J. van den Heuvel, R. J. M. M. Does, and J. P. S. Verver. "Six Sigma in Healthcare: Lessons Learned from a Hospital". In: *Int. J. Six Sigma Compet. Advant.* 1.4 (Dec. 2005), pp. 380–388. doi: 10.1504/IJSSCA.2005.008504.

[18] O. McDermott, J. Antony, S. Bhat, R. Jayaraman, A. Rosa, G. Marolla, and R. Parida. "Lean Six Sigma in Healthcare: A Systematic Literature Review on Motivations and Benefits". In: *Process.* 10.10 (Oct. 2022), p. 1910. doi: 10.3390/pr10101910.

[19] E. Bottani, B. Bigliardi, and B. Franchi. "Process Optimization in the Hospital Environment: A Systematic Review of the Literature and Results' Analysis". In: *Procedia Comput. Sci.* 200 (2022), pp. 1674–1684. doi: 10.1016/j.procs.2022.01.368.

[20] R. Rathi, A. Vakharia, and M. Shadab. "Lean Six Sigma in the Healthcare Sector: A Systematic Literature Review". In: *Int. Conf. Funct. Mater. Manuf. Perform.* Elsevier, Sept. 2021, pp. 773–781. doi: 10.1016/j.matpr.2021.05.534.

[21] F. Lalys and P. Jannin. "Surgical Process Modelling: a Review". In: *Int. J. Comput. Assist. Radiol. Surg.* 9 (May 2014), pp. 495–511. doi: 10.1007/s11548-013-0940-5.

[22] K. N. Timoh, A. Huaulme, K. Cleary, M. A. Zaheer, V. Lavoué, D. Donoho, and P. Jannin. "A Systematic Review of Annotation for Surgical Process Model Analysis in Minimally Invasive Surgery based on Video". In: *Surg. Endosc.* 37 (May 2023), pp. 4298–4314. doi: 10.1007/s00464-023-10041-w.

[23] P. Hartzband and J. Groopman. “Medical Taylorism”. In: *New Engl. J. Med.* 374.2 (Jan. 2016), pp. 106–108. doi: 10.1056/NEJMp1512402.

[24] S. Winch and A. J. Henderson. “Making Cars and Making Health Care: A Critical Review”. In: *Med. J. Aust.* 191.1 (July 2009), pp. 28–29. doi: 10.5694/j.1326-5377.2009.tb02670.x.

[25] I. Pernek and A. Ferscha. “A Survey of Context Recognition in Surgery”. In: *Med. Biol. Eng. Comput.* 55.10 (Oct. 2017), pp. 1719–1734. doi: 10.1007/s11517-017-1670-6.

[26] K. C. Demir, H. Schieber, T. Weise, D. Roth, M. May, and A. Maier. “Deep Learning in Surgical Workflow Analysis: A Review of Phase and Step Recognition”. In: *IEEE J. Biomed. Health Inform.* 27.11 (Nov. 2023), pp. 5405–5417. doi: 10.1109/JBHI.2023.3311628.

[27] A. Amin, S. A. Cardoso, J. Suyambu, H. A. Saboor, R. P. Cardoso, A. Husnain, N. V. Isaac, H. Backing, D. Mehmood, M. Mehmood, and A. N. J. Maslamani. “Future of Artificial Intelligence in Surgery: A Narrative Review”. In: *Cureus* 16.1 (Jan. 2024). doi: 10.7759/cureus.51631.

[28] Mayo Clinic Staff. *Coronary Angiogram*. Dec. 2023. url: https://www.mayoclinic.org/tests-procedures/coronary-angiogram/about/pac-20384904.

[29] R. Beyar, J. Davies, C. Cook, D. Dudek, P. Cummins, and N. Bruining. “Robotics, Imaging, and Artificial Intelligence in the Catheterisation Laboratory”. In: *EuroIntervention* 17.7 (Sept. 2021), pp. 537–549. doi: 10.4244/EIJ-D-21-00145.

[30] M. G. Bourassa. “The History of Cardiac Catheterization”. In: *Can. J. Cardiol.* 21.12 (Oct. 2005), pp. 1011–1014.

[31] W. Forssmann. “Die Sondierung des Rechten Herzens”. In: *Klin. Wochenschr.* 8 (Nov. 1929), pp. 2085–2087. doi: 10.1007/BF01875120.

[32] T. Barbour-Taylor, L. Mueller, D. Paris, D. Weaver, S. Amick, P. Bartzak, A. B. Britt, B. Brown, J. Bush, D. E. King, J. Richter, M. G. Webb, and A. Wood. *Pharmacology for Nurses*. Rice University, May 2024. url: https://openstax.org/books/pharmacology/.

[33] S. M. Levin. “The First Cardiac Catheter”. In: *J. Vasc. Surg.* 59.6 (June 2014), pp. 1744–1746. doi: 10.1016/j.jvs.2012.06.086.

[34] J. B. West. “The Beginnings of Cardiac Catheterization and the Resulting Impact on Pulmonary Medicine”. In: *Am. J. Physiol.* 313.4 (Oct. 2017), pp. L651–L658. doi: 10.1152/ajplung.00133.2017.

[35] A. V. G. Bruschke, W. C. Sheldon, E. K. Shirey, and W. L. Proudfit. “A Half Century of Selective Coronary Arteriography”. In: *J. Am. Coll. Cardiol.* 54.23 (Dec. 2009), pp. 2139–2144. doi: 10.1016/j.jacc.2009.06.051.

[36] F. M. Sones Jr. and E. K. Shirey. “Cine Coronary Arteriography”. In: *Mod. Concepts Cardiovasc. Dis.* 31 (1962), pp. 735–738.

[37] R. Bajaj, R. Parasa, A. Ramasamy, N. Makariou, N. Foin, F. Prati, A. Lansky, A. Mathur, A. Baumbach, and C. V. Bourantas. “Computerized Technologies Informing Cardiac Catheterization and Guiding Coronary Intervention”. In: *Am. Heart J.* 240 (Oct. 2021), pp. 28–45. doi: 10.1016/j.ahj.2021.05.017.

[38] B. H. W. Hendriks, D. Mioni, W. Crooijmans, and H. van Houten. “Image-Guided Intervention and Therapy: The First Time Right”. In: *Future Trends in Microelectronics: Journey into the Unknown*. Wiley, Sept. 2016. Chap. 3.1, pp. 243–258. doi: 10.1002/9781119069225.ch3-1.

[39] A. Roguin, P. Wu, T. Cohoon, F. Gul, G. Nasr, N. Premyodhin, and M. J. Kern. “Update on Radiation Safety in the Cath Lab – Moving Toward a “Lead-Free” Environment”. In: *Journal of the Society for Cardiovascular Angiography & Interventions* 2.4 (July 2023), p. 101040. doi: 10.1016/j.jscai.2023.101040.

[40] IBM. *What is AI?* Aug. 2024. url: https://www.ibm.com/think/topics/artificial-intelligence.

[41] S. Dong, P. Wang, and K. Abbas. “A Survey on Deep Learning and its Applications”. In: *Comput. Sci. Rev.* 40 (May 2021), p. 100379. doi: 10.1016/j.cosrev.2021.100379.

[42] W. J. Dally, S. W. Keckler, and D. B. Kirk. “Evolution of the Graphics Processing Unit (GPU)”. In: *IEEE Micro* 41.6 (Nov. 2021), pp. 42–51. doi: 10.1109/MM.2021.3113475.

[43] I. H. Sarker. “Machine Learning: Algorithms, Real-World Applications and Research Directions”. In: *SN Comput. Sci.* 2.3 (May 2021), p. 160. doi: 10.1007/s42979-021-00592-x.

[44] K. K. Parhi and N. K. Unnikrishnan. “Brain-Inspired Computing: Models and Architectures”. In: *IEEE Open J. Circuits Syst.* 1 (2020), pp. 185–204. doi: 10.1109/OJCAS.2020.3032092.

[45] W. S. McCulloch and W. Pitts. “A Logical Calculus of the Ideas Immanent in Nervous Activity”. In: *Bull. Math. Biophys.* 5 (Dec. 1943), pp. 115–133. doi: 10.1007/BF02478259.

[46] F. Rosenblatt. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”. In: *Psychol. Rev.* 65.6 (Nov. 1958), pp. 386–408. doi: 10.1037/h0042519.

[47] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri. “Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark”. In: *Neurocomputing* 503 (Sept. 2022), pp. 92–108. doi: 10.1016/j.neucom.2022.06.111.

[48] F. Rosenblatt. “Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms”. In: Spartan Books, 1962. Chap. Part III - Multi-Layer and Cross-Coupled Perceptrons, pp. 313–468.

[49] K. Hornik, M. Stinchcombe, and H. White. “Multilayer Feedforward Networks are Universal Approximators”. In: *Neural Netw.* 2.5 (1989), pp. 359–366. doi: 10.1016/0893-6080(89)90020-8.

[50] Q. Wang, Y. Ma, K. Zhao, and Y. Tian. “A Comprehensive Survey of Loss Functions in Machine Learning”. In: *Ann. Data Sci.* 9.2 (Apr. 2022), pp. 187–212. doi: 10.1007/s40745-020-00253-5.

[51] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. “Learning Representations by Back-propagating Errors”. In: *Nat.* 323.6088 (Oct. 1986), pp. 533–536. doi: 10.1038/323533a0.

[52] O. F. Razzouki, A. Charroud, Z. E. Allali, A. Chetouani, and N. Aslimani. “A Survey of Advanced Gradient Methods in Machine Learning”. In: *Int. Conf. Adv. Commun. Technol. Netw.* IEEE, Dec. 2024. doi: 10.1109/CommNet63022.2024.10793249.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Int. Conf. Neural Inf. Process. Syst.* Curran Associates Inc., Dec. 2012, pp. 1097–1105. doi: 10.5555/2999134.2999257.

[54] N. Padoy. “Machine and Deep Learning for Workflow Recognition during Surgery”. In: *Minim. Invasive Ther. Allied Technol.* 28.2 (Mar. 2019), pp. 82–90. doi: 10.1080/13645706.2019.1584116.

[55] V. Srivastav, T. Issenhuth, K. Abdolrahim, M. de Mathelin, A. Gangi, and N. Padoy. “MVOR: A Multi-View RGB-D Operating Room Dataset for 2D and 3D Human Pose Estimation”. In: *Conf. Med. Image Comput. Comput. Assist. Interv.* MICCAI, 2018.

[56] V. Belagiannis, X. Wang, H. B. B. Shitrit, K. Hashimoto, R. Stauder, Y. Aoki, M. Kranzfelder, A. Schneider, P. Fua, S. Ilic, H. Feussner, and N. Navab. “Parsing Human Skeletons in an Operating Room”. In: *Mach. Vis. Appl.* 27 (Oct. 2016), pp. 1035–1046. doi: 10.1007/s00138-016-0792-4.

[57] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin, and N. Padoy. “Articulated Clinician Detection using 3D Pictorial Structures on RGB-D Data”. In: *Med. Image Anal.* 35 (Jan. 2017), pp. 215–224. doi: 10.1016/j.media.2016.07.001.

[58] A. Kadkhodamohammadi and N. Padoy. “A Generalizable Approach for Multi-View 3D Human Pose Regression”. In: *Mach. Vis. Appl.* 32 (Oct. 2021), p. 6. doi: 10.1007/s00138-020-01120-2.

[59] V. F. Rodrigues, R. S. Antunes, L. A. Seewald, R. Bazo, E. S. dos Reis, U. J. L. dos Santos, R. d. R. Righi, L. G. d. S. Junior, C. A. da Costa, F. L. Bertollo, A. Maier, B. Eskofier, T. Horz, M. Pfister, and R. Fahrig. “A Multi-Sensor Architecture combining Human Pose Estimation and Real-Time Location Systems for Workflow Monitoring on Hybrid Operating Suites”. In: *Future Gener. Comput. Syst.* 135 (Oct. 2022), pp. 283–298. doi: 10.1016/j.future.2022.05.006.

[60] K. Yokoyama, G. Yamamoto, C. Liu, K. Kishimoto, and T. Kuroda. “Operating Room Surveillance Video Analysis for Group Activity Recognition”. In: *Adv. Biomedic. Eng.* 12 (2023), pp. 171–181. doi: 10.14326/abe.12.171.

[61] K. Yokoyama, G. Yamamoto, C. Liu, O. Sugiyama, L. H. Santos, and T. Kuroda. “Recognition of Instrument Passing and Group Attention for Understanding Intraoperative State of Surgical Team”. In: *Adv. Biomedic. Eng.* 11 (2022), pp. 37–47. doi: 10.14326/abe.11.37.

I

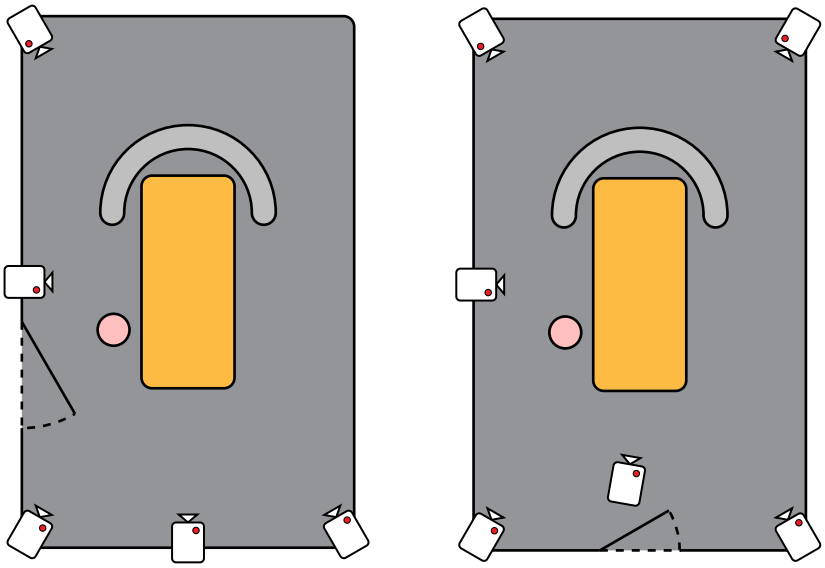
FEATURE EXTRACTION


2

COMPUTER VISION IN THE CARDIAC CATHETERISATION LABORATORY

Workflow insights can improve efficiency and safety in the Cardiac Catheterization Laboratory (Cath Lab). As manual analysis is labor-intensive, we aim for automation through camera monitoring. Automated workflow analysis from monitoring footage requires computers to 'see' into the room. This chapter explores several computer vision techniques to recognise objects and persons in cardiac catheterisation laboratories and operating rooms. Besides the detection methods, generalisation to different rooms and viewpoints is considered. A pipeline is proposed that uses camera calibrations to detect persons in three-dimensional space. Most objects were detected with an average precision above 0.9, although small objects yielded scores down to 0.72. When training and testing on different datasets, all average precision scores dropped, depending on the room and object. Cameras were calibrated with a Euclidean distance error of 32.8 mm. 2D human poses were triangulated to 3D with a Euclidean distance error of at most 76.5 mm. The proposed pipeline yields promising results, although it generalises poorly. For scalable applications, personnel and the display were most reliably detected accross interventional rooms. Calibration and triangulation worked well, but required a new annotated dataset for each new room. Therefore, for scalability, 2D human poses seem to be the most reliable feature to work with.

Parts of this chapter were published in Symp. Inf. Theory Signal Process. Benelux. (2022), © WIC [1],
Int. Conf. Acoust. Speech, Signal Process. (2023), © IEEE [2], and
IEEE Int. Workshop Med. Meas. Appl. (2024), © IEEE [3].



(a) Reinier (b) Philips

Figure 2.1.: Top-down recording setups in two Cath Labs. Camera positions and entrances are shown. The operating table, C-arm and cardiologist are drawn schematically for reference.

Public hospitals have been coping with increasing operational costs and personnel shortages for years [4, 5]. Resulting stress on the healthcare system increases patient waiting times, personnel workload and, as a result, operational costs for the hospital. Operating Rooms (ORs) represent a major expense for hospitals to operate. Methods to decrease costs and workload are an active topic of research. One approach in literature is to focus on procedure- and turnover efficiency [6]. Identifying and teaching good workflow practices could improve efficiency and cost-effectiveness whilst maintaining or decreasing workload [7, 8].

The identification of best intraoperative workflow practices requires extensive observations to be made during procedures. To get good coverage of workflow practices, one needs to make observations for different surgical teams or in multiple hospitals. The large amount of manual labour this entails calls for a more easily applicable solution that is generalisable. Automatic analysis through computer algorithms could achieve this.

Workflow analysis algorithms need datastreams from the OR that contain workflow information. Examples are measurements from sensors, medical imaging, and patient records [9]. None of the above provide information during the entire procedure. An explorative solution

is to use computer vision (CV) in a monitoring setup. Cameras can be mounted in the OR, recording procedures from a distance. CV can extract descriptive features from the images, providing input for further processing or analysis.

Neural Networks (NNs) have proven to be an effective tool for image analysis. Convolutional NNs have kickstarted a revolution in CV [10]. Low-level image features are extracted and combined into higher-level features over a number of steps. The resulting highest-level features are used to make a prediction, based on the specific task the NN was designed to do. A NN is trained to extract descriptive features based on its task by showing it examples from an annotated dataset like [11]. After training, when shown an unseen image, a well-trained NN is often able to predict a fitting annotation. The NN model structure, dataset, and annotations dictate the task that it will be able to perform.

ORs present unique visual challenges such as clutter, reflections, and occlusion during procedures. Due to privacy regulations, such environments are scarcely represented in public datasets. At the time of writing, the only public dataset with real surgeries we are aware of is MVOR [12]. It is not guaranteed that algorithms developed for more general situations still perform well in the OR.

Reference [13] investigates 3D human pose estimation from multi-view OR videos by directly optimising a 3D loss function, yielding a strong dependence on the camera system. This non-generalisability is partially solved in [14], which performs 2D- and 3D pose regression separately. The NNs that this method employs for 3D regression, though, need retraining for different camera systems and is therefore not generalisable. Reference [15] implements unsupervised domain adaptation by imposing geometric constraints, maintaining the need to retrain but removing the requirement of new annotations. All these methods assume that the camera system is calibrated, whereas in practice cameras are often moved in ORs for various reasons.

This work summarises four research projects on (generalisable) object detection, automated camera calibration, and human keypoint detection in a clinical setting [1–3, 16–19]. Object detection and human keypoint detection provide information about human actions and person-person or person-object interactions during procedures. Calibration allows detections from multiple synchronised viewpoints to be triangulated to coordinates in three-dimensional space, mitigating the occlusion that individual viewpoints suffer from. Automated periodic calibration makes the method robust to cameras being moved, triangulating 3D poses is generalisable to arbitrary calibrated camera systems without retraining. This pilot takes place in several cardiac catheterisation laboratories (Cath Labs): specialised ORs that are equipped for minimally invasive interventional cardiac procedures [20]. Monitoring cameras were hung in Cath Labs, as visualised in fig. 2.1. The combined result of the

considered research projects is an algorithm that produces object- and human pose locations in the Cath Lab, triangulated using automated camera calibration [1].

2.1. COMPUTER VISION TASKS

This section presents a set of object detection tasks and challenges studied in literature. Sections 2.1.1 to 2.1.3 list methods to extract information from images and video, which usually employ NNs. Section 2.1.4 discusses the deployment of such methods in new environments, that were not seen during training. The relation of locations in images to different viewpoints or real-world coordinates is discussed in sections 2.1.5 and 2.1.6.

2.1.1. OBJECT DETECTION

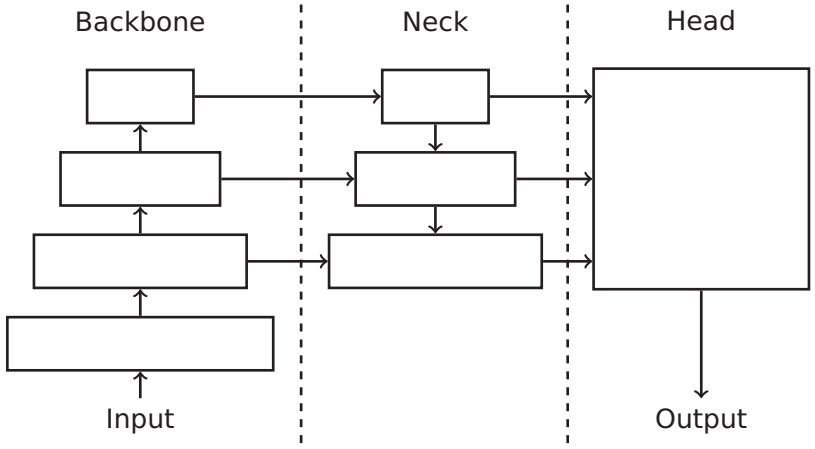


Figure 2.2.: Architecture of a neural network for object detection. The backbone extract features from input images at multiple levels of detail. The neck aggregates these features from multiple levels. Finally, the head produces bounding box predictions from the aggregated features. In this example, the backbone consists of four neural network layers, the neck of three, and the head can have any architecture.

An object detector aims to find the bounding box and class of all objects in an image. Many modern object detectors are structured in three consecutive parts [21], as visualised in fig. 2.2:

- The backbone extracts features from an input image, gradually producing higher-level features in a sequence of processing steps.
- The neck aggregates low- and high-level backbone features.

- The head predicts object bounding boxes and classes from the neck output.

There are two approaches to the operation of the head. Two-stage detectors first localise object bounding boxes, then classify each. One-stage detectors perform localisation and classification simultaneously. Often, one-stage detectors are less accurate but faster, making them more suitable for real-time applications.

A well-known one-stage object detector is You Only Look Once (YOLO) [22]. This algorithm is a Convolutional Neural Network (CNN) that divides an input image into a grid and predicts an object class and a number of bounding boxes per grid cell. YOLO evolved through the years, attracting increasing attention and making rapid speed- and accuracy improvements [23].

CNNs use sliding windows to extract or refine feature maps, and have dominated the field of CV. Later, Transformers were proposed [24], which employ a self-attention mechanism to analyse relations within a sequence. Representing an image as a sequence of smaller images yielding the Vision Transformer (ViT) [25] for use in CV. In contrast to CNNs, ViTs encode global relations and suppress features irrelevant to the problem. Self-attention has been used in some iterations of YOLO [23].

2.1.2. POSE DETECTION

Physical object states like e.g. orientation or pose cannot be deduced from bounding boxes. Pose- or landmark detection aims te detect a set of pre-defined keypoints in objects [26]. For instance, [11] defines seventeen keypoints that form a human pose: the nose, eyes, ears, shoulders, elbows, wrists, hips, knees and ankles. We consider multi-object pose detection, which makes no assumptions on the number of objects in view. There are two common approaches to this:

- Top-down estimators apply an object detector to obtain object bounding boxes. Within each bounding box, a heatmap is generated for each keypoint class. Keypoints are regressed at the maximum of their respective heatmap.
- Bottom-up estimators generate heatmaps over an entire image. Different heatmaps encode e.g. the probability of keypoint class presence, or information about present keypoint relations. Information of all heatmaps is combined to place keypoints and connect them into object poses.

Like object detectors, most pose detectors work on features extracted with a backbone. These are refined to maps that encode e.g. the probability that a keypoint or keypoint relation is present. Decoding

such maps yields a set of objects and their poses. Many pose detectors are CNNs, although Transformers have been applied as well.

2.1.3. MULTI-OBJECT TRACKING

Objects and poses are detected in still images. In video, the same object may appear on several frames. To analyse e.g. motion, it is necessary to reidentify (ReID) these objects. Each detection is assigned an identifier (ID). If the same object is recognised on another frame, it is assigned the same ID. This process is called multi-object tracking (MOT) [27]. In this work, we refer to groups of detections with the same ID as 'tracklets'.

Many MOT approaches are CNN-based [27]. Objects can be reidentified using any combination of:

- Motion features to extrapolate the next object position and see if it overlaps a new detection.
- Visual features to evaluate if a detection visually resembles a previously seen object.
- Temporal features to analyse visual changes between frames.

Occlusion and sterile clothing complicate the use of visual features for MOT in the Cath Lab.

2.1.4. DOMAIN SHIFT

Machine Learning algorithms trained in one environment may not work in another [28, 29]. This 'domain shift' manifests in various ways [30, 31], such as covariate shift: when input data are not distributed the same when training and applying the algorithm. Cath Labs show covariate shift, e.g., when the same object looks different in two Cath Labs, lighting is different, or different camera angles- or systems are used [3].

Domain shift can be overcome with domain adaptation- or generalisation [28, 29]. Domain adaptation uses data from a new environment to adapt an algorithm trained in another. Domain generalisation strives for domain-invariance by extracting general—rather than domain-specific—features:

- Data manipulation augments or generates data to simulate many domains during training.
- Representation learning separates features into domain-specific and general components, e.g., by comparing features from multiple domains (domain alignment).

- Learning strategy teaches an algorithm general knowledge, e.g., by optimising an optimiser itself for this purpose (meta-learning) [32].

2.1.5. CAMERA CALIBRATION

Camera calibration is the process of finding camera properties that define how it records the world [33]. Camera parameters can be divided into two groups:

- Intrinsics: the focal length and sensor offset relative to the lens, both inherent to the camera.
- Extrinsics: The position and orientation of the camera with respect to a global coordinate system.

Intrinsic parameters can be found using known calibration patterns. Extrinsics can be estimated given the intrinsics, and a set of 2D image- and 3D real-world coordinate correspondences. This is called the Perspective-n-Point (PnP) problem. If cameras can move, it is necessary to recalibrate extrinsics periodically. Placing permanent calibration patterns in a clinical setting is not practical, so other reference points should be used.

2.1.6. CAMERA GEOMETRY

Camera calibration provides a mapping between its 2D image coordinates and 3D world coordinates. For instance, a point in the camera image can be mapped onto a straight line in the real world [34]. If the same point is seen from multiple calibrated cameras, its real-world position can be inferred from such lines using triangulation. Also, a 2D point in one camera view can be projected onto another as an 'epipolar line'.

In this work, we start by explaining the object detection, calibration and human pose detection methods in section 2.2. Evaluation metrics are described in the same section. section 2.3 shows the evaluation outcomes, and their implications are discussed in section 2.4. Finally, section 2.5 concludes the work.

2.2. PROPOSED PIPELINE

This section starts with an explanation of the recorded dataset in section 2.2.1. Each remaining subsection explains a different part of the image processing pipeline shown in fig. 2.3.

2.2.1. DATASETS & ANNOTATIONS

Various datasets were used to train and evaluate CV algorithms.

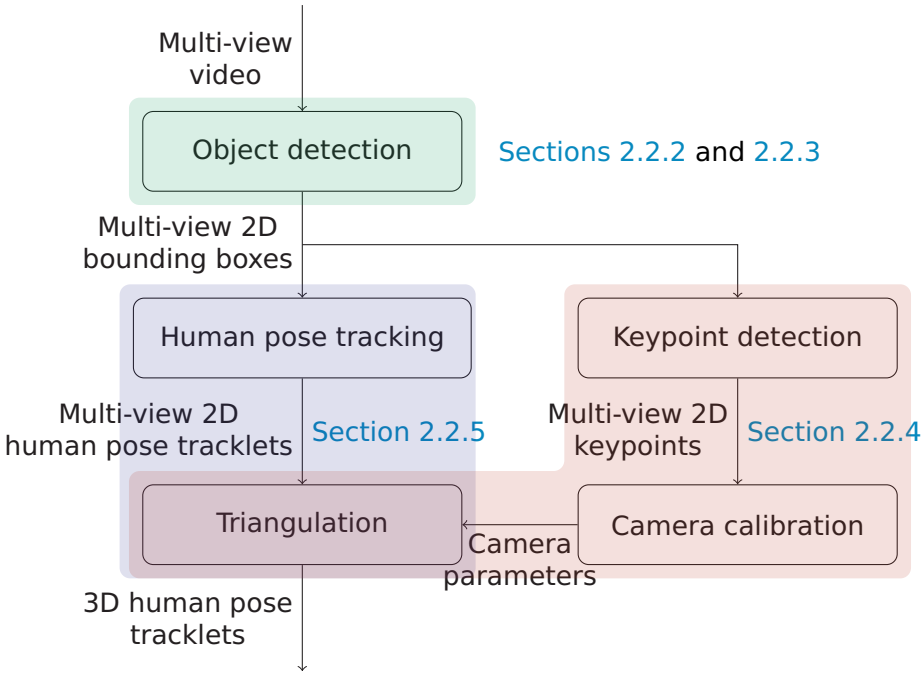


Figure 2.3.: Pipeline for object bounding box detection, camera calibration, and 3D human pose detection.

REINIER

Table 2.1.: The number of annotated bounding boxes in videos of the empty Reinier de Graaf Cath Lab and interventions, taken from [18].

Object	Empty room	Interventions
Cardiologist	0	2869
Assistant	961	7959
Patient	0	5578
Instrument table	434	5373
Operating table	800	7690
Control panel display	0	9075
Control panel buttons	0	7841
X-ray detector	818	6488
X-ray source	304	4540
Lead shield	484	10377
Radiation light	0	905
Monitor	991	4982

A dataset was recorded in the Cath Lab of the Reinier de Graaf Gasthuis, Delft, NL, with four Axis M1125 and one Axis M3046-V ceiling-mounted cameras. The recording setup is visualised in fig. 2.1a, where the camera on the left wall is the M3046-V. The cameras record video in a resolution of 1920 px $\times$ 1088 px at 25 frames per second (fps), but were downsampled to 0.2 fps for annotation. Bounding boxes of Cath Lab-specific objects were annotated by the authors of [18, 19] from all viewpoints during three videos: one of the empty Cath Lab, and two during real cardiac angiograms procedures. The video of the empty Cath Lab contained 991 frames, and the remaining videos together 6218 frames. The total number of annotated bounding boxes is shown in table 2.1.

REINIER MOBILE

Another dataset was recorded in [2, 17] in the same Cath Lab, with a mobile phone. Around each camera in the Reinier dataset, pictures were taken of the room from sixteen angles. Bounding boxes, 2D image keypoints, and corresponding 3D real-world keypoints were annotated of fixed objects: the doors, windows, dresser, monitor, and a stretcher hanging on the wall. The annotated keypoints were placed on object corners and vertices.

PHILIPS

One of the companies developing Cath Labs is Philips Healthcare, Best, NL. Two mock procedures were recorded in a Cath Lab there, where the roles of cardiologist, assistants and patient were played by Philips

personnel. The recordings were shorter than those in the Reinier dataset, and recorded with 25 fps. The same classes were annotated as in the Reinier dataset by the author of [19], with the exception that no distinction was made between the cardiologist and assistants. The recording setup is shown in fig. 2.1b.

ONLINE

Cath Lab images are available online, mostly for advertising purposes. They are usually recorded from a low viewpoint, contrasting with the ceiling-mounted cameras of the previously described datasets. Wide-view cameras are often used to capture the complete Cath Lab, introducing distortion. Online images show a wide range of different Cath Labs with varying backgrounds and object appearances. A dataset was composed in [19] by searching for 'Catheterization Laboratory' on Google Images and Bing Images. Images that were blurry, damaged, irrelevant to the query, or that had a resolution below 640 px $\times$ 480 px were excluded. The resulting dataset was annotated like the previous two sets. Compared to the Reinier and Philips sets, this online dataset shows almost no cardiologists, assistants and patients.

OPERATING ROOM

Reference [35] published a dataset of mock procedures in an operating room. Videos were stored at a resolution of 1280 px $\times$ 720 px with a framerate of 1 fps. The cameras were calibrated using a checkerboard pattern on the floor. Human poses were annotated with nine landmarks: the head, torso, shoulders, elbows, wrists, and stomach. No legs were annotated, as these were poorly visible due to occlusion from the room and sterile clothing.

2.2.2. MULTI-VIEW OBJECT DETECTION

To detect the objects from table 2.1, Scaled-YOLOv4-CSP [42] was trained on the Reinier dataset in [18]. Scaled-YOLOv4 is an iteration of YOLO [22] that was made scalable to enable a tradeoff between speed and performance. Scaled-YOLOv4-CSP is one of its scaled versions, balancing performance and speed. At the end of the backbone, a multi-head self-attention layer [24, 25] was added to encode global relations between regions of the feature map.

Due to occlusion, objects are not always clearly visible from all viewpoints. Therefore, [18] refines object detections in a target viewpoint using information from neighbouring viewpoints. This process is illustrated in fig. 2.5. First, SuperPoint [39] detects keypoints where geometric information is dense in the target- and neighbouring views. Per bounding box in the neighbouring views, SuperGlue [40] matches

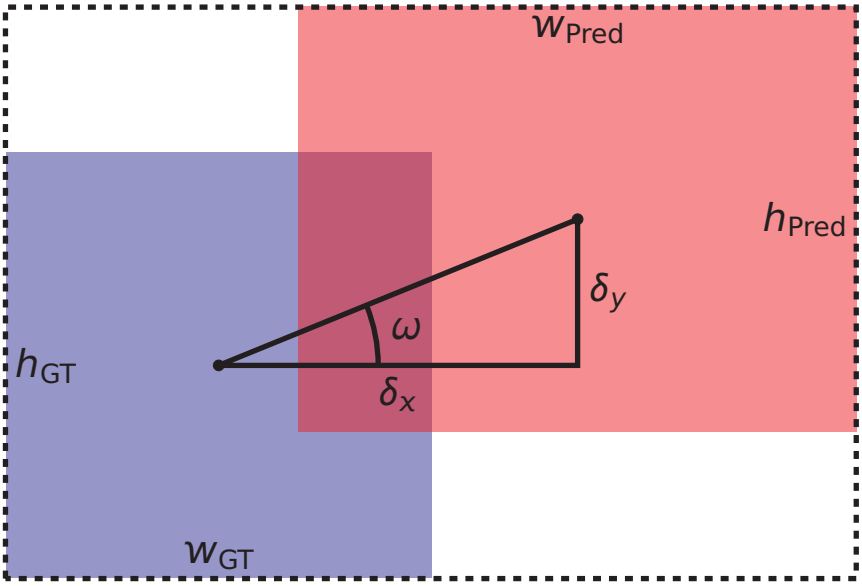
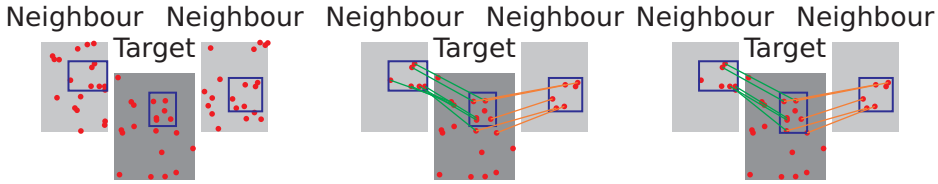


Figure 2.4.: Example ground-truth (GT) and predicted (Pred) bounding boxes. IoU loss, which aims to maximise IoU, knows several extensions. Generalised-IoU loss [36] minimises the smallest rectangle encompassing both bounding boxes. Complete-IoU loss [37] minimises centre distance $\sqrt{\delta_x^2 + \delta_y^2}$ and the difference between aspect ratios $w_{\text{GT}}/h_{\text{GT}}$ and $w_{\text{Pred}}/h_{\text{Pred}}$. SCYLLA-IoU loss [38] aligns the centres orthogonally ($\omega \in \{0, \pi/2\}$), minimises orthogonal centre distances δ_x and δ_y, and minimises dimension differences between $\begin{bmatrix} w_{GT} & h_{GT} \end{bmatrix}$ and $\begin{bmatrix} w_{Pred} & h_{Pred} \end{bmatrix}$.



(a) A frame from a target- and two neighbouring viewpoints. In each viewpoint, keypoint detections from SuperPoint [39] and an object detection are shown.

(b) Keypoints in the neighbouring views are filtered by the object bounding box. Keypoints are matched from each neighbouring viewpoint to the target viewpoint with SuperGlue [40].

(c) Background keypoint matches are removed with DBSCAN [41]. The bounding box in the target viewpoint is expanded to fit all remaining matched keypoints.

Figure 2.5.: Bounding box refinement using multi-view information.

the keypoints in this bounding box to any keypoints in the target view. This yields a list of target view keypoints that correspond to an object in the neighbouring view. This list may contain matched keypoints of background objects, which were filtered out using the DBSCAN clustering algorithm [41]. The bounding box can then be expanded such that all listed keypoints lie inside it. Although perspective differences cause each neighbour to miss object keypoints visible in the target view, this is compensated by using both neighbours. Because of the camera setup, the northwestern and southeastern views each have only one neighbour.

Scaled-YOLOv4-CSP and the multi-head self-attention layer were initialised randomly.The model was trained on two videos from the Reinier dataset that show the empty Cath Lab and one procedure, using SCYLLA-Intersection over Union (SIoU) loss [38]. fig. 2.4 shows the operation of SIoU, which aims to align the predicted- and ground truth bounding box centres, minimise differences between their dimensions, and maximise Intersection over Union (IoU). Detection performance was evaluated with the average precision (AP) metric. Testing was done on the single Reinier procedure that was not used for training.

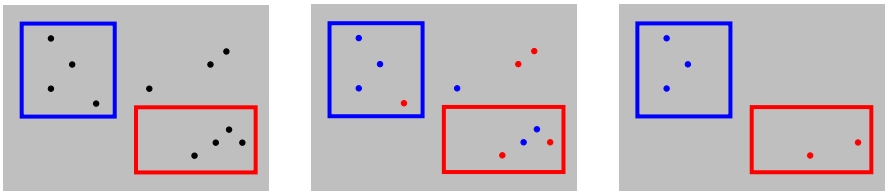
2.2.3. OBJECT DETECTION WITH DOMAIN SHIFT

Reference [3] investigates domain shift- and generalisation in Cath Labs using YOLOv8 [43]. YOLOv8 uses multiple heads that separately perform classification and bounding box regression. To visualise domain shift, features were extracted from all datasets using weights[1] pretrained on

[1]Available: https://github.com/ultralytics/ultralytics/blob/main/docs/en/models/yolov8.md

the COCO dataset. The resulting features were cast to two dimensions using UMAP [44], giving one datapoint per image. To test generalisation, YOLOv8 was finetuned on the Online dataset and evaluated on the remaining data. This was done using Complete-IoU (CIoU) loss [37]. As visualised in fig. 2.5, CIoU loss aims to maximise IoU, and minimise prediction-ground truth centre distance and aspect ratio difference. During this work, the 'Cardiologist' and 'Lab assistant' were merged into a 'Staff' class, and the lead shield and radiation light were left out for consistency between datasets.

2.2.4. CAMERA CALIBRATION



(a) Two objects and multiple keypoint detections.

(b) Assigning keypoint classes based on matching to Reinier Mobile.

(c) Keypoints outside the bounding box of their class are excluded.

Figure 2.6.: Excluding keypoints based on object detection results for more accurate camera calibration.

The Reinier dataset cameras were calibrated in [2, 17]. Since calibration should be non-intrusive, i.e., require no permanent calibration pattern placement in Cath Labs, object- and keypoint detectors were used to find reference points.

For object detection, Scaled-YOLOv4-P5 was trained on the Reinier dataset using Generalised-IoU (GIoU) loss [36]. Rather than the classes of table 2.1, training was done on the fixed objects from Reinier Mobile. The model was initialised with pre-trained weights[2] obtained using the COCO dataset [11]. Besides maximising IoU, GIoU loss minimises the smallest rectangle encompassing the prediction and ground truth.

Keypoints were detected in camera images with SuperPoint [39]. These were matched to annotated keypoints in the Reinier Mobile dataset using SuperGlue [40]. Matching was done with each of the sixteen Reinier Mobile images corresponding to the calibrating camera view. Thus, each keypoint detection was matched to at most sixteen keypoint annotations. As corresponding 2D and 3D keypoint coordinates were annotated in Reinier Mobile, the result was a set of correspondences between the camera image and real-world coordinates.

[2]Available: https://github.com/WongKinYiu/ScaledYOLOv4/

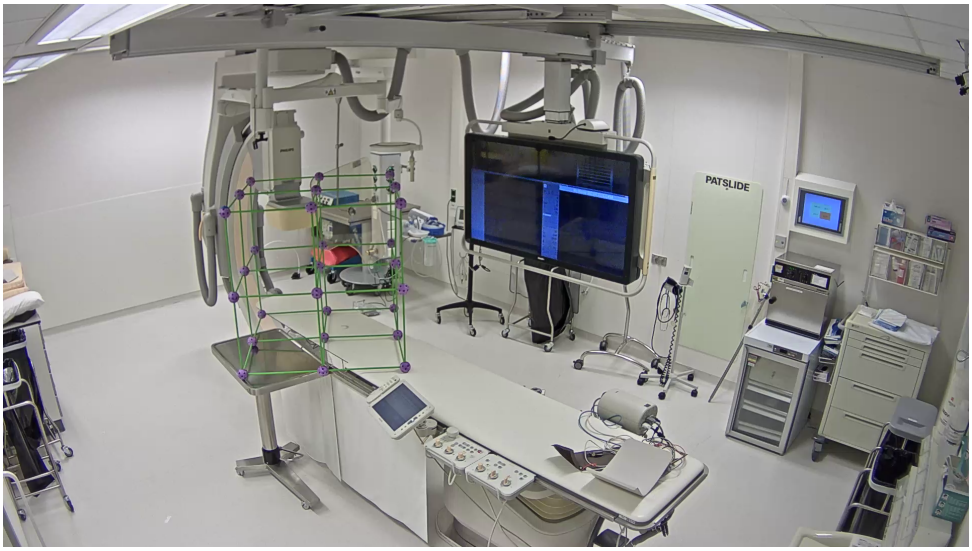


Figure 2.7.: Wireframe of known dimensions to evaluate calibration metrics.

Keypoint matches were filtered using bounding box detections, as visualised in fig. 2.6. Each keypoint match was assigned a class label based on the annotated keypoint. Any matches which detected keypoint were not inside a bounding box of the same class, were discarded.

Keypoint detections may be matched to multiple keypoint annotations and, consequently, associated with multiple 3D coordinates. In such cases, the mean 3D coordinate was calculated. The annotated 3D coordinate lying closest to this mean was selected, and the other matches discarded. The EPnP algorithm [45] was used to estimate extrinsics from the remaining matches between camera image- and real-world coordinates.

For evaluation purposes, the wireframe shown in fig. 2.7 was placed in the Cath Lab. Its 27 vertices were annotated from each Reinier viewpoint from fig. 2.1a. The calibration was used to find the 3D coordinate of each vertex using Linear-Eigen triangulation [34]. Then, the triangulated vertices were reprojected onto each camera image to obtain a new set of 2D vertex coordinates. Reprojection error (RPE) measured the distance between the annotated- and reprojected 2D vertices in px. Euclidean distance error (EDE) evaluated the distance between the triangulated and actual 3D vertices in mm. Both RPE and EDE were averaged over the camera viewpoints and wireframe vertices.

2.2.5. HUMAN POSE ESTIMATION

Reference [16] explores 3D human pose tracking in the Cath Lab. First, 2D person bounding boxes are detected in each viewpoint of the Reinier dataset using YOLOX-x [46]. YOLOX-x imposes no prior on bounding box regression, decouples classification and regression into separate heads, and compensates for the case where ground-truth bounding boxes overlap. A model[3] pretrained on the COCO dataset [11] was finetuned on the Operating Room dataset. As only humans needed to be detected, the final layer was reduced to have only one output channel. Bounding boxes were tracked using ByteTrack[4] [47], which associates bounding boxes based on detection confidence, appearance, and predicted linear motion.

In each bounding box, a human pose was regressed using HRNet-W48 [48]. This top-down pose estimator maintains high-resolution features for precise keypoint placement. Like YOLOX-x, the pretrained model[5] on COCO [11] was finetuned on the Operating Room dataset. The number of output channels was reduced, as the Operating Room dataset defines different pose landmarks than COCO.

3D RECONSTRUCTION

3D poses can be reconstructed from multi-view 2D poses, if it is known which detections show the same individual. Detections were matched similarly as in [49], which considers all viewpoints simultaneously: Given a set of D detections over all viewpoints $\mathcal{D} = \{d \mid 1 \leq d \leq D, d \in \mathbb{N}\}$, each entry $r_{n,m}$ in affinity matrix

$$R = \begin{bmatrix} r_{1,1} & \cdots & r_{1,D} \\ \vdots & \ddots & \vdots \\ r_{D,1} & \cdots & r_{D,D} \end{bmatrix} \in \mathbb{R}^{D \times D} \tag{2.1}$$

scores the probability that $n \in \mathcal{D}$ and $m \in \mathcal{D}$ show the same person.

Scores $r_{n,m}$ consisted of two parts. Epipolar affinity measured how close each pose is to the epipolar projection of the other pose onto its viewpoint:

$$r^{e}_{n,m} = \text{sigmoid}\left(\frac{1}{2\,|\mathcal{K}|} \sum_{K \in \mathcal{K}} \delta\left(d^{(f)}_{n,K}, e_{j \to m}(d^{(f)}_{m,K})\right) + \delta\left(d^{(f)}_{m,K}, e_{n \to m}(d^{(f)}_{n,K})\right) \right), \tag{2.2}$$

where $\mathcal{K}$ is the set of predefined keypoints, $\delta(p, l)$ measures the distance between point p and line l, $d^{(f)}_{p,K}$ is the detection of keypoint class K in

[3] Available: https://github.com/Megvii-BaseDetection/YOLOX
[4] Available: https://github.com/ifzhang/ByteTrack
[5] Available: https://github.com/HRNet/HRNet-Human-Pose-Estimation

pose p on frame f, and $e_{m\to n}(d_{m,K}^{(f)})$ is the epipolar line of $d_{m,K}^{(f)}$ in the viewpoint of detection n. The sigmoid function casts $r_{n,m}^{e}$ to $(0,1)$. If n and m were in the same viewpoint, $r_{n,m}^{e}$ was set to zero.

Tracking affinity evaluated if the same objects were matched in the previous frame:

$$r_{n,m}^{t} = \begin{cases} 1 & \text{if } d_n^{(f-1)} \text{ and } d_m^{(f-1)} \text{ were matched} \\ 0 & \text{otherwise} \end{cases}, \tag{2.3}$$

where $d_n^{(f-1)}$ and $d_m^{(f-1)}$ are full-pose detections from the tracklets of n and m on the previous frame. Unlike in [49] no visual features were used, as sterile clothing in the Cath Lab complicates matching from visuals. The final affinity score was

$$r_{n,m} = \sqrt{\zeta r_{n,m}^{e} + (1-\zeta) r_{n,m}^{t}}, \tag{2.4}$$

where $\zeta \in [0,1]$ is a weight.

A permutation matrix $Q \in \mathbb{R}^{D\times D}$ was found from R with entries

$$q_{n,m} = \begin{cases} 1 & \text{if } n \text{ and } m \text{ are matched} \\ 0 & \text{otherwise} \end{cases}. \tag{2.5}$$

The construction of Q was formulated as a convex linear problem:

$$\underset{Q}{\text{minimise}} \; -\langle R, Q\rangle + \lambda \|Q\|_*, \tag{2.6}$$

s.t.

$$q_{n,m} = q_{m,n} \quad \forall\, n \in \mathcal{D}, m \in \mathcal{D}, \tag{2.7}$$

$$0 \le q_{n,m} \le 1 \quad \forall\, n \in \mathcal{D}, m \in \mathcal{D}, \tag{2.8}$$

$$0 \le Q\mathbf{1} \le 1, \tag{2.9}$$

$$0 \le Q^T\mathbf{1} \le 1, \tag{2.10}$$

where $\langle \square \rangle$ denotes the Frobenius inner product, λ weights the second part of the objective function, and $\|\square\|_*$ is the nuclear norm, which approaches matrix rank in a convex and continuous way. Equation (2.7) ensures that n is matched to m if and only if m is matched to n. Equations (2.8) to (2.10) relax the constraints of a permutation matrix to allow non-discrete optimisation. The problem was solved with the alternating direction method of multipliers [50]. After convergence, Q is discretised with thresholding.

For each match between the five viewpoints, the top three most confident human poses were linearly triangulated [34] to construct a 3D pose. As no pose annotations existed for the Reinier dataset, results were evaluated on the Operating Room dataset. Reconstruction was evaluated by calculating the EDE in mm between 3D landmark

detections, and ground-truth 2D landmarks that were mapped onto 3D lines. Similarly, EDE was calculated in px between the 2D ground truths and reprojected 3D poses. Both metrics were averaged over the five viewpoints and detected poses.

2.3. RESULTS

2.3.1. MULTI-VIEW OBJECT DETECTION

Table 2.2.: Average precision from [18] of trained Scaled-YOLOv4-CSP on the Reinier dataset. The model was trained and evaluated by itself, with an added multi-head self-attention layer, and with refinement by neighbouring viewpoints.

Label	AP@.5:.95 Scaled-YOLOv4-CSP	+ multi-head self-attention	+ multi-view refinement
Cardiologist	0.946	0.940	0.963
Lab assistant	0.866	0.886	0.914
Patient	0.957	0.968	0.959
Instrument table	0.966	0.971	0.948
Operating table	0.966	0.975	0.967
Control panel display	0.964	0.971	0.868
Control panel buttons	0.934	0.931	0.846
X-ray detector	0.975	0.978	0.939
X-ray source	0.983	0.988	0.923
Lead shield	0.964	0.971	0.937
Radiation light	0.763	0.836	0.717
Display	0.995	0.995	0.994

Object detection results from [18] are shown in table 2.2. Scaled-YOLOv4-CSP already detects most classes with an AP above 0.9. Exceptions are the Lab assistant at 0.866 and Radiation light at 0.763.

Adding a multi-head self-attention layer improved AP for must classes with up to 0.02. Results rose further for the Radiation light, with 0.073. The multi-head self-attention layer lowered AP for the Cardiologist and Control panel buttons with 0.006 and 0.003.

Bounding box refinement with information from neighbouring viewpoints lowered AP for most classes, by up to 0.085. AP for the control panel display and Radiation light deteriorated further, by 0.103 and 0.119.

2.3.2. OBJECT DETECTION WITH DOMAIN SHIFT

Dimension-reduced features from different Cath Labs and viewpoints are shown in fig. 2.8. Datapoints of the same dataset and camera gathered

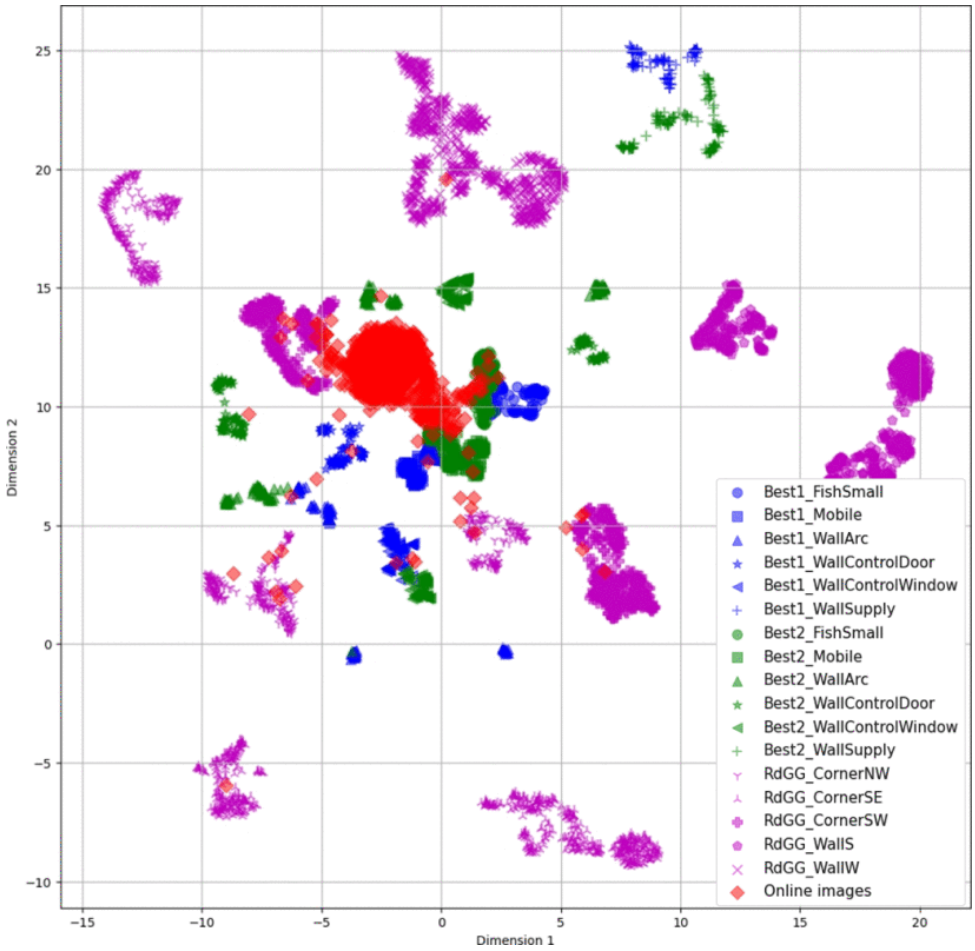


Figure 2.8.: Video data distributions in the Reinier, Philips, and Online datasets, visualised using UMAP [44]. Cath Lab dataset are shown in different colours, and viewpoints with different markers. The two Philips procedures are dubbed ‘Best1’ and ‘Best2’, and the Reinier dataset ‘RdGG’. This figure was taken from [3, 19].

Table 2.3.: Average precision from [19] on several datasets, after training on the Online dataset.

Label	AP@.5 Reinier procedure 1	Philips procedure 1	Philips procedure 2
Staff	0.856	0.870	0.831
Patient	0.274	0.485	0.516
Instrument table	0.704	0.649	0.581
Operating table	0.266	0.616	0.670
Control panel display	0.738	0.152	0.214
Control panel buttons	0.311	0.115	0.054
X-ray detector	0.709	0.684	0.766
X-ray source	0.071	0.681	0.630
Display	0.727	0.974	1.000

into clusters, although multiple clusters existed for most cameras. Clusters lied further apart within the Reinier dataset than the Philips and Online datasets. The Online dataset is spread out the least, and has most overlap with other datasets. The Reinier and Philips datasets barely overlap.

Table 2.3 shows evaluated AP@.5 on several datasets, after training on the Online dataset. Staff and the display were detected consistently with AP scores of at least 0.727. For the patient and control panel buttons, no AP above 0.516 was ever achieved. The operating table and X-ray source were detected with at least 0.616 AP during both Philips procedures, but only 0.266 and 0.071 AP in the Reinier dataset. These roles are reversed for the control panel display, which is detected with 0.738 AP in the Reinier dataset and at most 0.214 AP at Philips.

2.3.3. CAMERA CALIBRATION

Table 2.4.: Mean reprojection error and Euclidean distance error from [2, 17] after calibrating the five cameras of the Reinier dataset. Results are shown both with and without bounding box-based match exclusion.

Match exclusion	RPE (px)	EDE (mm)
No	56.83	87.7
Yes	19.95	32.8

RPE and EDE after calibration of the Reinier dataset cameras are shown in table 2.4. Results are shown both when considering all keypoint matches, and after excluding keypoints using bounding box detections. Keypoint exclusion improves RPE and EDE by 36.88 px and

54.9 mm respectively. The average RPE of 19.95 px amounts to 1.04 % of the image width.

2.3.4. HUMAN POSE ESTIMATION

Table 2.5.: Human pose Euclidean distance error from [16] on the Operating Room dataset. metrics were calculated between 2D human pose annotations and 3D pose detections, in image- and real-world coordinates, and reported per landmark.

Landmark	EDE (px)	EDE (mm)
Head	15.80	54.13
Torso	13.84	50.98
Left shoulder	13.59	54.05
Right shoulder	13.73	53.16
Left elbow	16.05	65.62
Right elbow	16.26	66.56
Left wrist	17.76	76.53
Right wrist	18.06	76.54
Stomach	15.35	68.04
Average	15.64	62.84

EDE between ground-truth 2D human poses and 3D pose detections, evaluated on the Operating Room dataset, can be seen in table 2.5. EDE varies between 13.59 px and 18.06 px in image space and between 50.98 mm and 76.54 mm in world coordinates. On both metrics, the error was largest for the wrists and lowest for the torso and shoulders.

2.4. DISCUSSION

In this work, we summarised and related four research projects on CV in the Cath Lab. A pipeline was built to obtain 2D object bounding boxes, camera calibrations and 3D human poses from multi-view monitoring videos. Several datasets were introduced and used for algorithm development and testing.

Object detection was accurate when training and testing in the Reinier Cath Lab. Nearly all objects—even the transparent lead shield—were detected with a high AP@.5:.95 above 0.914. Exceptions were the radiation light and control panel. YOLO object detectors are known to struggle with such small objects, although this has improved in some iterations [23]. The addition of a self-attention layer slightly increased

AP for most classes. Bounding box refinement using neighbouring views did not yield improvements, except for detection of the cardiologist and lab assistants. This could be due to the limitation that only bounding boxes that were already detected were refined. Using neighbouring views to correct false negatives- or positives may have a larger impact, as occlusion is a problem for single-view object detection in Cath Labs.

Training and testing in different Cath Labs dropped performance significantly. AP varied between Cath Labs, suggesting that room- and object appearances introduced a large domain shift. One major issue was that the control panel and X-ray device were covered in plastic during procedures, which was not the case in the Online dataset. In addition, the viewpoints and camera intrinsics were different in the Online dataset than in the Reinier- and Philips Cath Labs. Online images were relatively close to each other in feature space, whereas during procedures big differences were present between rooms and viewpoints. Object detection could be generalised in Cath Labs with a larger training dataset that covers more of the feature space, or representation learning to bring features closer together.

After camera calibration, keypoints in the centre of the room could be localised within 3.28 cm EDE on average. This seems sufficient to measure overall position, excluding subtle movements of e.g. the fingers. Keypoint match filtering proved a large improvement, more than halving EDE. A current limitation is the necessity of additional recordings around each viewpoint, with landmark annotations and measurements. An alternative could be matching keypoints between viewpoints for calibration [51]. This approach would yield a linear transformation of real-world camera coordinates, as no reference pattern is used. Viewpoint sparsity may present a problem for this solution.

Detected 2D human poses were succesfully triangulated to world coordinates. The EDE was higher than measured during calibration. This was expected, as EDE during calibration was calculated using only annotated landmarks. For pose detection, faults in 2D pose estimation and matching between views will have increased average EDE. EDE was higher for the wrists, elbows and stomach than other landmarks. As many catheterisation actions are carried out using the arms, wrist movements are likely to contain workflow information. This means it would be worthwile to try and improve EDE for the arms.

An average 3D pose EDE of 62.8 mm was measured whereas [14] achieved 11 mm, although this was on a different dataset. There is no question that their method was better optimised, with 3D pose regression done using a NN trained on annotated images. Here, there is a tradeoff between the generalisability of our pipeline and the accuracy of theirs.

Each research project trained their own object detector to detect specific classes. To build a full, efficient pipeline, a single detector could

be trained to detect everything. Keypoint detections- and matches from object detection refinement could be reused during calibration. Pose estimation was the only algorithm that was tested in an operating room instead of a Cath Lab.

All pipeline blocks except the 2D human pose tracking relied on Cath Lab-specific object detection. As demonstrated, data distributions are not the same between Cath Labs. Without good object detection, the proposed calibration is not possible. As mentioned this could be solved by calibrating without landmark detection, although this introduces other limitations. For the pipeline to generalise to other Cath Labs, e.g. representation learning should be investigated in the object detection. The integration of multiple views mitigates occlusion and adds redundancy, yielding more robust results in Cath Labs.

3D objects and human poses may act as input to further analysis. For instance, they may be used for activity recognition [52] or workflow phase recognition [8]. Understanding occurrences in the room may automate e.g. summary writing after procedures, or enable (partial) autonomy of some devices like the C-arm. Activity timelines could be presented in dashboards, or compared to optimise workflow practices or object placement.

2.5. CONCLUSION

The proposed computer vision pipeline performs robust object detection, camera calibration and 3D pose estimation in Cath Labs. As retraining is currently required for each new Cath Lab, generalisation needs to be improved to obtain a scalable solution.

REFERENCES

[1] Y. Jiang, R. Dai, J. Zeng, R. Butler, T. Vijfvinkel, Y. Wang, J. J. van den Dobbelsteen, M. van der Elst, and J. Dauwels. “Object Detection and Person Tracking in CathLab with Automatically Calibrated Cameras”. In: *Symp. Inf. Theory Signal Process. Benelux*. WIC, June 2022, p. 57. url: https://www.w-i-c.org/proceedings/proceedings_SITB2022.pdf.

[2] J. Zeng, R. Butler, J. J. van den Dobbelsteen, B. H. W. Hendriks, M. van der Elst, and J. Dauwels. “Automatic Camera Pose Estimation by Key-Point Matching of Reference Objects”. In: *Int. Conf. Acoust., Speech, Signal Process.* IEEE, June 2023. doi: 10.1109/ICASSP49357.2023.10095197.

[3] Z. Wang, R. Butler, J. J. van den Dobbelsteen, B. H. W. Hendriks, M. van der Elst, and J. Dauwels. “Towards Robust Object Detection in Unseen Catheterization Laboratories”. In: *IEEE Int. Workshop Med. Meas. Appl.* IEEE, June 2024. doi: 10.1109/MeMeA60663.2024.10596906.

[4] R. Marjamaa, A. Vakkuri, and O. Kirvelä. “Operating Room Management: Why, How and by Whom?” In: *Acta Anaesthesiol. Scand.* 52.5 (Apr. 2008), pp. 596–600. doi: 10.1111/j.1399-6576.2008.01618.x.

[5] C. B. E. Halbeis and A. Schubert. “Staffing the Operating Room Suite: Perspectives from Europe and North America on the Role of Different Anesthesia Personnel”. In: *Anesthesiol. Clinic.* 26.4 (Dec. 2008), pp. 637–663. doi: 10.1016/j.anclin.2008.07.002.

[6] A. Pasquer, S. Ducarroz, J. C. Lifante, S. Skinner, G. Poncet, and A. Duclos. “Operating Room Organization and Surgical Performance: a Systematic Review”. In: *Patient Saf. Surg.* 18.1 (Jan. 2024), p. 5. doi: 10.1186/s13037-023-00388-3.

[7] C. von Schudnat, K.-P. Schoeneberg, J. Albors-Garrigos, B. Lahmann, and M. De-Miguel-Molina. “The Economic Impact of Standardization and Digitalization in the Operating Room: A Systematic Literature Review”. In: *J. Med. Syst.* 47 (Dec. 2023), p. 55. doi: 10.1007/s10916-023-01945-0.

[8] A. M. Schouten, S. M. Flipse, K. E. van Nieuwenhuizen, F. W. Jansen, A. C. van der Eijk, and J. J. van den Dobbelsteen. “Operating Room Performance Optimization Metrics: a Systematic Review”. In: *J. Med. Syst.* 47.1 (Dec. 2023), p. 19. doi: 10.1007/s10916-023-01912-9.

[9] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, M. Hashizume, D. Katic, H. Kenngott, M. Kranzfelder, A. Malpani, K. März, T. Neumuth, N. Padoy, C. Pugh, N. Schoch, D. Stoyanov, R. Taylor, M. Wagner, G. D. Hager, and P. Jannin. “Surgical Data Science for Next-generation Interventions”. In: *Nat. Biomed. Eng.* 1 (Sept. 2017), pp. 691–696. doi: 10.1038/s41551-017-0132-7.

[10] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parnar. “A Review of Convolutional Neural Networks in Computer Vision”. In: *Artif. Intell. Rev.* 57.4 (Apr. 2024), p. 99. doi: 10.1007/s10462-024-10721-6.

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. “Microsoft COCO: Common Objects in Context”. In: *Eur. Conf. Comput. Vis.* Springer, Sept. 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[12] V. Srivastav, T. Issenhuth, K. Abdolrahim, M. de Mathelin, A. Gangi, and N. Padoy. “MVOR: A Multi-View RGB-D Operating Room Dataset for 2D and 3D Human Pose Estimation”. In: *Conf. Med. Image Comput. Comput. Assist. Interv.* MICCAI, 2018.

[13] B. G. A. Gerats, J. M. Wolterink, and I. A. M. J. Broeders. “3D Human Pose Estimation in Multi-view Operating Room Videos using Differentiable Camera Projections”. In: *Comput. Methods Biomech. Biomed. Eng.: Imaging Vis.* 11.4 (2023), pp. 1197–1205. doi: 10.1080/21681163.2022.2155580.

[14] A. Kadkhodamohammadi and N. Padoy. “A Generalizable Approach for Multi-View 3D Human Pose Regression”. In: *Mach. Vis. Appl.* 32 (Oct. 2021), p. 6. doi: 10.1007/s00138-020-01120-2.

[15] V. Srivastav, A. Gangi, and N. Padoy. “Unsupervised Domain Adaptation for Clinician Pose Estimation and Instance Segmentation in the Operating Room”. In: *Med. Image Anal.* 80 (Aug. 2022), p. 102525. doi: 10.1016/j.media.2022.102525.

[16] Y. Jiang. “Automated Personnel Activities Observation in the Catheterization Laboratory”. MA thesis. Delft University of Technology, July 2022. url: http://resolver.tudelft.nl/uuid:ec5ac611-9524-4e45-b8ce-b9b1dbfe2fe7.

[17] J. Zeng. “Automatic Camera Extrinsics Estimation in the Catheterization Laboratory”. MA thesis. Delft University of Technology, Sept. 2022. url: http://resolver.tudelft.nl/uuid:91f7878b-ef71-47db-90b5-d4b48cefb314.

[18] R. Dai. “Detecting Medical Equipment in the Catheterization Laboratory using Computer Vision”. MA thesis. Delft University of Technology, Sept. 2022. url: http://resolver.tudelft.nl/uuid:1894799d-302b-44bc-841e-bf652471330b.

[19] Z. Wang. “Towards Robust Object Detection in Unseen Catheterization Laboratories”. MA thesis. Delft University of Technology, Jan. 2024. url: http://resolver.tudelft.nl/uuid:c4b2d25c-0b2e-41e1-9d2b-32b9e0c88bb2.

[20] Mayo Clinic Staff. *Cardiac Catheterization*. Aug. 2023. url: https://www.mayoclinic.org/tests-procedures/cardiac-catheterization/about/pac-20384695.

[21] S. Bouraya and A. Belangour. “Deep Learning based Neck Models for Object Detection: A Review and a Benchmarking Study”. In: *Int. J. Adv. Comput. Sci. Appl.* 12.11 (Nov. 2021), p. 19. doi: 10.14569/IJACSA.2021.0121119.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. “You Only Look Once: Unified, Real-Time Object Detection”. In: *Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[23] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González. “A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS”. In: *Mach. Learn. Knowl. Extr.* 5.4 (Dec. 2023), pp. 1680–1716. doi: 10.3390/make5040083.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is all you Need”. In: *Int. Conf. Neural Inf. Process. Syst.* Curran Associates Inc., Dec. 2017, pp. 6000–6010. url: https://dl.acm.org/doi/10.5555/3295222.3295349.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *Int. Conf. Learn. Represent.* May 2021, p. 7. url: https://openreview.net/forum?id=YicbFdNTTy.

[26] H. Chen, R. Feng, S. Wu, H. Xu, F. Zhou, and Z. Liu. “2D Human Pose Estimation: A Survey”. In: *Multimed. Syst.* 29 (Oct. 2023), pp. 3115–3138. doi: 10.1007/s00530-022-01019-0.

[27] C. Du, C. Lin, R. Jin, B. Chai, Y. Yao, and S. Su. “Exploring the State-of-the-Art in Multi-Object Tracking: A Comprehensive Survey, Evaluation, Challenges, and Future Directions”. In: *Multimed. Tool. Appl.* 83 (Sept. 2024), pp. 73151–73189. doi: 10.1007/s11042-023-17983-2.

[28] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. “Domain Generalization: A Survey”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 45.4 (Apr. 2023), pp. 4396–4415. doi: 10.1109/TPAMI.2022.3195549.

[29] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, and W. Lu. “Generalization to Unseen Domains: A Survey on Domain Generalization”. In: *IEEE Trans. Knowl. Data Eng.* 35.8 (Aug. 2023), pp. 8052–8072. doi: 10.1109/TKDE.2022.3178128.

[30] A. Torralba and A. A. Efros. “Unbiased Look at Dataset Bias”. In: *Conf. Comput. Vis. Pattern Recognit.* IEEE, Aug. 2011, pp. 1521–1528. doi: 10.1109/CVPR.2011.5995347.

[31] R. Alaiz-Rodríguez and N. Japkowicz. “Assessing the Impact of Changing Environments on Classifier Performance”. In: *Conf. Can. Soc. Comput. Stud. Intell.* Springer, May 2008, pp. 13–24. doi: 10.1007/978-3-540-68825-9_2.

[32] T. Hospedales, A. Antoniou, and P. Micaelli. “Meta-Learning in Neural Networks: A Survey”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 33.9 (Sept. 2022), pp. 5149–5169. doi: 10.1109/TPAMI.2021.3079209.

[33] J. Salvi, X. Armangué, and J. Batlle. “A Comparative Review of Camera Calibrating Methods with Accuracy Evaluation”. In: *Pattern Recognit.* 35.7 (July 2002), pp. 1617–1635. doi: 10.1016/S0031-3203(01)00126-1.

[34] R. I. Hartley and P. Sturm. “Triangulation”. In: *Comput. Vis. Image Underst.* 68.2 (Nov. 1997), pp. 146–157. doi: 10.1006/cviu.1997.0547.

[35] V. Belagiannis, X. Wang, H. B. B. Shitrit, K. Hashimoto, R. Stauder, Y. Aoki, M. Kranzfelder, A. Schneider, P. Fua, S. Ilic, H. Feussner, and N. Navab. “Parsing Human Skeletons in an Operating Room”. In: *Mach. Vis. Appl.* 27 (Oct. 2016), pp. 1035–1046. doi: 10.1007/s00138-016-0792-4.

[36] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. “Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression”. In: *Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2019, pp. 658–666. doi: 10.1109/CVPR.2019.00075.

[37] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, and Q. Hu. “Enhancing Geometric Factors in Machine Learning and Inference for Object Detection and Instance Segmentation”. In: *IEEE Trans. Cybern.* 52.8 (Aug. 2022), pp. 8574–8586. doi: 10.1109/TCYB.2021.3095305.

[38] Z. Gevorgyan. *SIoU Loss: More Powerful Learning for Bounding Box Regression*. May 2022. doi: 10.48550/arXiv.2205.12740. arXiv: 2205.12740 [cs.CV].

[39] D. DeTone, T. Malisiewicz, and A. Rabinovich. “SuperPoint: Self-Supervised Interest Point Detection and Description”. In: *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*. IEEE, June 2018, pp. 337–349. doi: 10.1109/CVPRW.2018.00060.

[40] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. “SuperGlue: Learning Feature Matching with Graph Neural Networks”. In: *Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2020, pp. 4937–4946. doi: 10.1109/CVPR42600.2020.00499.

[41] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. “A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *Proc. Second Int. Conf. Knowl. Discov. Data Min.* AAAI Press, Aug. 1996, pp. 226–231. url: https://dl.acm.org/doi/10.5555/3001460.3001507.

[42] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. “Scaled-YOLOv4: Scaling Cross Stage Partial Network”. In: *Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2021, pp. 13024–13033. doi: 10.1109/CVPR46437.2021.01283.

[43] R. Varghese and M. Sambath. “YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness”. In: *Int. Conf. Adv. Data Eng. Intell. Comput. Syst.* IEEE, Apr. 2024. doi: 10.1109/ADICS58448.2024.10533619.

[44] L. McInnes, J. Healy, N. Saul, and L. Großberger. “UMAP: Uniform Manifold Approximation and Projection”. In: *J. Open Source Softw.* 3.29 (Sept. 2018), p. 861. doi: 10.21105/joss.00861.

[45] V. Lepetit, F. Moreno-Noguer, and P. Fua. “EPnP: An Accurate O(n) Solution to the PnP Problem”. In: *Int. J. Comput. Vis.* 81 (Feb. 2009), pp. 155–166. doi: 10.1007/s11263-008-0152-6.

[46] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. *YOLOX: Exceeding YOLO Series in 2021*. Aug. 2021. doi: 10.48550/arXiv.2107.08430. arXiv: 2107.08430 [cs.CV].

[47] Y. Zhang, P. Sun, y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang. “ByteTrack: Multi-object Tracking by Associating Every Detection Box”. In: *Eur. Conf. Comput. Vis.* Springer, Oct. 2022, pp. 1–21. doi: 10.1007/978-3-031-20047-2_1.

[48] K. Sun, B. Xiao, D. Liu, and J. Wang. “Deep High-Resolution Representation Learning for Human Pose Estimation”. In: *Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2019, pp. 5686–5696. doi: 10.1109/CVPR.2019.00584.

[49] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou. “Fast and Robust Multi-person 3D Pose Estimation from Multiple Views”. In: *Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2019, pp. 7784–7793. doi: 10.1109/CVPR.2019.00798.

[50] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers”. In: *Found. Trends Mach. Learn.* 3.1 (2011), pp. 1–122. doi: 10.1561/2200000016.

[51] D. Nister. “An Efficient Solution to the Five-Point Relative Pose Problem”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 26.6 (June 2004), pp. 756–770. doi: 10.1109/TPAMI.2004.17.

[52] G. Saleem, U. I. Bajwa, and R. H. Raza. “Toward Human Activity Recognition: a Survey”. In: *Neural Comput. Appl.* 35 (Feb. 2023), pp. 4145–4182. doi: 10.1007/s00521-022-07937-4.

3

BENCHMARKING 2D HUMAN POSE TRACKERS

Literature shows that human poses are indicative of activities and therefore workflow. As an exploration, we evaluate how marker-less multi-human pose estimators perform in the Cath Lab. Poses were annotated in 2040 frames from ten multi-view coronary angiogram (CAG) recordings. Pose estimators AlphaPose, OpenPifPaf and OpenPose were run on the footage. Detection and tracking were evaluated separately for the Head, Arms, and Legs with Average Precision (AP), head-guided Percentage of Correct Keypoints (PCKh), Association Accuracy (AA), and Higher-Order Tracking Accuracy (HOTA). We give qualitative examples of results for situations common in the Cath Lab, with reflections in the monitor or occlusion of personnel. AlphaPose performed best on most mean Full-pose metrics with an AP from 0.56 to 0.82, AA from 0.55 to 0.71, and HOTA from 0.58 to 0.73. On PCKh OpenPifPaf scored highest, from 0.53 to 0.64. Arms, Legs, and the Head were detected best in that order, from the views which see the least occlusion. During tracking in the Cath Lab, AlphaPose tended to swap identities and OpenPifPaf merged different individuals. Results suggest that AlphaPose yields the most accurate confidence scores and limbs, and OpenPifPaf more accurate keypoint locations in the Cath Lab. Occlusions and reflection complicate pose tracking. The AP of up to 0.82 suggests that AlphaPose is a suitable pose detector for workflow analysis in the Cath Lab, whereas its HOTA of up to 0.73 here calls for another tracking solution.

Parts of this chapter were published in Med. Eng. Phys. **136** (2025), © Elsevier [1].

The field of workflow analysis is gaining traction in medical environments [2–5]. During surgery, insight into workflow is necessary in order to optimise procedures. Example use-cases are improved procedure efficiency, safety, and training.

Manual workflow analysis is a laborious task that requires experts to carry out. Automation enables cost-effective, large-scale deployment and additional use-cases like real-time feedback or support [6–9]. Personnel activities, which can be found from human pose tracklets [10–12], are descriptive of workflow.

Multi-object keypoint detection—also called pose estimation—aims to localise predefined objects and their keypoints in an image. Figure 3.2 shows keypoints and edges ('limbs') for the 'Human' class as defined in [13]. Pose estimators output a continuous pixel (px) location and confidence score per detected keypoint. Modern works often take one of two approaches:

- Top-down: Detect object bounding boxes [14] and estimate a pose in each of them [15–17].
- Bottom-up: Detect keypoints and assemble them into objects [18–20].

In this work we refer to human keypoints using the abbreviations and groupings from fig. 3.1, where a leading '*l*' or '*r*' denotes 'left' or 'right'.

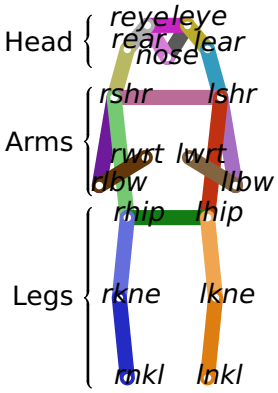


Figure 3.1.: COCO pose [13] facing the reader

The temporal element in videos gives need to multi-object tracking: the assignment of the same identity (ID) to the same object in different video frames. This can be done causally [21], non-causally [22], jointly with detection [19], or separately after detection [15]. We denote algorithms with tracking capabilities with a superscript '*T*'.

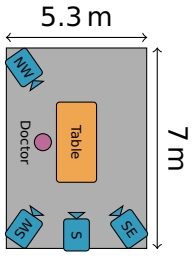
Annotations are required to train or test keypoint detectors and trackers. Human annotators label 'ground-truth' poses with their

(a) CornerNW

(b) CornerSE

(c) CornerSW

(d) WallS



(e) Cath Lab dimensions in meters

Figure 3.2.: (a)-(d) Camera viewpoints with annotated poses, (e) Map of the Cath Lab with measurements in meters, table and cameras not to scale.

video-specific ID, presence, and location. Algorithms learn to mimic the annotation process and are tested against ground-truths.

Medical environments present challenges like significant occlusion between personnel and objects, and appearance similarities due to sterile clothing. General-purpose datasets like [13] are not representative of such settings. Evaluating pose estimator performance requires recordings of real procedures, which are scarce due to privacy regulations [23]. MVOR [24] is a public dataset with recordings from the hybrid operating room (OR). It was recorded in four days during different procedures in a university hospital. To capture workflow information, however, more data and procedure uniformity are needed.

Human pose estimation in ORs was investigated in [25–27]. Reference [27] tests a state-of-the-art 2D pose estimator on a single metric, and focusses on the step to 3D pose estimation. To our knowledge, optimality of the chosen 2D pose estimator in a medical setting has not been verified. Reference [28] investigates the scalability of object detection to different Cath Labs, but does not consider pose detection.

The Cardiac Catheterization Laboratory (Cath Lab) is a specialised operating room (OR) where minimally invasive cardiovascular procedures take place. This work evaluates the performance of human pose estimators and trackers as a potential tool for workflow analysis in the Cath Lab. To this end, we record real coronary angiogram (CAG) procedures in a regional hospital from the four camera (Axis M1125) views shown in fig. 3.2. The videos capture workflow before, during and after procedures. Poses are annotated in ten procedures, showing five

different workflow phases. The Cath Lab presents unique challenges to computer vision like concealing clothing, occlusion, and reflections. To our knowledge, no video dataset of real Cath Lab procedures exists in literature at the time of this study. An estimator to analyze anyfuture recordings can be selected in line with results from this work.

We test several pre-trained state-of-the-art 2D human pose estimators in the Cath Lab. Three algorithms were selected by the criteria that they i) can detect an arbitrary number of poses per image, and ii) provide implementation details in peer-reviewed work: AlphaPose [15] (AlphaP), OpenPifPaf [19] (OpenPP), and OpenPose [18] (OpenP). As AlphaP is a top-down estimator and OpenPP and OpenP are bottom-up, results should give an idea of which approach works best in the Cath Lab. AlphaP and OpenPP also provide causal tracking models AlphaP^T and OpenPP^T. We quantitatively measure detection- and tracking performance and support these metrics with qualitative observations.

The main contributions of this work are:

- We introduce a unique multi-view dataset of real CAG procedures in the Cath Lab with pose annotations.
- We evaluate the performance of several state-of-the-art 2D human pose estimators in the Cath Lab.
- We discuss—from a workflow perspective—pitfalls for pose estimation that are Cath Lab-specific.

Section 3.1 starts with a description of our dataset, included algorithms, and evaluated metrics. Section 3.2 lists the results and highlights trends and differences. Then, section 3.3 discusses and explains observed outcomes. We theorise what the results imply for our setting and identify an algorithm for use in future work. Finally, section 3.4 gives a summary.

3.1. MATERIALS AND METHODS

This section describes all components that make up our benchmark. Section 3.1.1 starts with a description of the dataset and its recording process. Section 3.1.2 provides a brief explanation of the used pose estimators. Section 3.1.3 concludes with our used metrics and other evaluations.

3.1.1. VIDEO RECORDINGS

Four cameras (Axis M1125) were hung in the Cath Lab of the Reinier de Graaf Gasthuis, Delft, NL. With approval of a local medical ethics committee and the hospital board, and informed consent from the patients and staff, CAG procedures were recorded from the viewpoints in

fig. 3.2 and stored with a resolution of 1920 px × 1088 px and framerate of 25 frames per second. A cardiologist, scrub nurse, up to two lab assistants, and the patient were present during each procedure. We record and annotate ten procedures, where we ensure that each shows a different medical team for variability. CAGs follow a strict, consistent workflow with little to no variation. Because of this uniformity, the ten chosen procedures cover the typical cases. Local doctors helped select the procedures to include some rare deviations. For instance, there is a procedure during which the cardiologist had to move the monitor, one where ultrasound was needed to find the radial artery for endovascular access, and one where the staff struggled to reposition the lead shield.

ANNOTATION

In each procedure, poses were annotated in 51 frames sampled uniformly over 30 seconds, from four synchronised viewpoints. This gives a total of 10 (procedures) × 51 (frames) × 4 (viewpoints) = 2040 annotated frames. The 30 seconds per procedure were hand-picked to show one of five unique workflow phases:

- The patient entering and lying down.
- Realization of endovascular access through the wrist.
- Use of ultrasound to detect the radial artery for endovascular access.
- X-Ray imaging.
- Closure of the entrywound.

Each phase was selected twice from different procedures. Poses were annotated in Computer Vision Annotation Tool (CVAT) [29] by two of the authors with a background in engineering, and their quality confirmed by a third who has been a practicing interventional cardiologist for over 13 years. We did not use the CVAT interpolation feature in order to preserve fine positioning, which we expect to be important for workflow analysis in the Cath Lab. One annotated example frame is shown per viewpoint in fig. 3.2. Fully occluded individuals and keypoint reflections in e.g. the monitor were not labelled. People in the control room and hallway were included.

We define a person to be ‘visible’ on a frame if any of their keypoints can be seen directly in that frame without obstruction. To describe the dataset we label each frame by presence of situations that arise in the Cath Lab:

- Occluded fully: A person is inside the camera view but not visible.
- Occluding person: Segmentations of visible persons overlap.

- Occluding object: An object segmentation overlaps a visible person.
- Occluding sheet: The surgical sheet overlaps the visible patient.
- Occluding clothes: Sterile clothes conceal visible elbows, knees or hips.
- Occluding window: The control room window overlaps a visible person.
- Occluding view: A wall or frame boundary overlaps a visible person.
- Horizontal patient: The visible patient shows non-vertically in the view.
- Reflecting monitor: The monitor shows a reflected person.
- Reflecting window: The control room window shows a reflected person.

Some situations are viewpoint-specific, e.g., CornerSE sees no reflective surfaces and the patient is vertical from WallS even when lying down. The situations are labelled per frame, i.e., if a situation occurs multiple times in the same frame it is counted as a single instance. In addition, we record the number of visible people per frame using the same methodology. Finally, we count the total number of annotated keypoints per class where multiple can be counted per frame.

3.1.2. POSE ESTIMATION

AlphaP is implemented as a parallel pipeline which aims for high inference speeds. A fast object detector [30, 31] detects Human bounding boxes, in each of which a Convolutional Neural Network (CNN) generates a heatmap per keypoint. At the maximum of this heatmap, the keypoint is placed. This per-bounding box processing makes AlphaP a top-down algorithm. Optionally a second CNN extracts features per bounding box for tracking and trajectory smoothing, where background noise is mitigated by masking with the detected pose. A low object detector confidence threshold avoids false negatives but yields redundant detections. Pose Non-Maximum Suppression removes resulting duplicate poses. A translation-invariant approximation of the loss function gradient is used during optimization. Additionally, heatmaps are normalised such that calculated confidences become invariant of keypoint scale.

OpenP has a CNN encode limb presence and orientation over the entire image into Part Affinity vector Fields (PAFs). A second CNN generates keypoint heatmaps and locations from these PAFs like AlphaP. Poses are assembled in bottom-up fashion with bipartite matching: each

candidate limb is scored by integration over its PAF, and a set of limbs is selected to maximise the sum of scores.

OpenPP replaces heatmaps with Composite Intensity Fields which encode keypoint confidence, scale, and location offset. PAFs are replaced with Composite Association Fields (CAFs) which encode i) probability of limb presence and ii) endpoint scales and location offsets. Temporal CAFs model limbs between keypoints of the same class in adjacent frames for tracking. Poses are grown bottom-up by greedy matching from a high-confidence seed keypoint, guided by these intensity and association fields. Keypoint-level Non-Maximum Suppression removes duplicate poses. Redundant limbs are modelled for robustness against occlusion.

3.1.3. EXPERIMENTAL SETUP

The following sections describe the used pre-trained models, evaluation metrics, and other validation procedures.

MODEL SETTINGS

Each algorithm offers several pre-trained models, which can be split into i) a backbone which extracts image features and ii) a head which estimates poses and/or IDs. We test the models in table 3.1 on our dataset without retraining, using an NVIDIA GeForce RTX 3090 GPU. The used model parameters are available publicly for AlphaP[1], OpenPP[2] and OpenP[3]. For fair comparison we sample all outputs to the format from fig. 3.1. We define pose confidence as the mean of all its nonzero keypoint confidences.

QUANTITATIVE METRICS

Metrics measure detection- and tracking performance per viewpoint. They are calculated with True Positives (TPs), False Positives (FPs), and False Negatives (FNs), using only visible keypoints.

Average Precision (AP) [13, 34] evaluates Full-pose detection. As it was designed for bounding boxes, we replace its use of Intersection over Union with Object Keypoint Similarity (OKS) as suggested in [13], where we estimate segmentation area with the tightest-fit pose bounding box. For a more detailed evaluation we calculate AP separately for three subposes: Head, Arms, and Legs, in addition to the Full pose. We calculate $AP^{\tau_{OKS}}$ at OKS thresholds $\tau_{OKS} = 0.5$ (low), $\tau_{OKS} = 0.75$ (high),

[1] Available: https://github.com/MVIG-SJTU/AlphaPose/blob/master/docs/MODEL_ZOO.md
[2] Available: https://openpifpaf.github.io/intro.html
[3] Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose/tree/master/models

Table 3.1.: Tested pose estimators where a superscript '*T*' denotes tracking capabilities.

Algorithm	Backbone	Head	Training dataset
AlphaP	YOLOv3-SPP[30, 31] +ResNet152[32]	FastPose (DUC)[15]	COCO[13]
AlphaP^T	YOLOv3-SPP[30, 31] +ResNet152[32]	FastPose (DUC) +Human-ReID[15]	COCO[13]
OpenPP	shufflenetv2k30[19, 33]	CifCaf[19]	COCO[13]
OpenPP^T	tshufflenetv2k30[19, 33]	TrackingPose[19]	COCO[13]
OpenP	OpenPose[18]	OpenPose[18]	COCO[13] +Human Foot[18]

and averaged from 0.5 to 0.95 with step size 0.05 $\tau_{\text{OKS}} = 0.5 : 0.95$ (ranged).

Head-guided Percentage of Correct Keypoints (PCKh) [35] evaluates detection per keypoint. We first use the Hungarian algorithm [36] to match annotated and estimated poses by OKS where—as opposed to AP—we do not threshold confidence. PCKh is evaluated per match. Since the COCO pose has no headbone, we threshold TPs with 0.5 times the longest annotated *shr-ear* distance instead, and only use poses with such an annotated limb. With the obtained per-keypoint TPs, FPs and FNs we calculate

$$\text{PCKh} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{3.1}$$

We evaluate tracking for each viewpoint and subpose with Association Accuracy (AA) [37], and replace its use of Localization Similarity with OKS as was done for AP. Finally, Higher-Order Tracking Accuracy (HOTA) [37] summarises detection and tracking performance in a single metric. Although HOTA is an aggregation of AA and Detection Accuracy, we do not evaluate the latter, as its purpose is similar to that of the more commonly used AP.

We show metrics evaluated per individual video, each of which shows one of five workflow phases from different procedures. Error bars show two standard deviations around the mean metric. If one situation yields a better score than others, we say this situation is 'preferred'. Unless explicitly stated otherwise, discussed results are mean Full-pose scores for ranged τ_{OKS}.

STATISTICAL SIGNIFICANCE

We evaluate the significance of performance differences between each pair of algorithms with a two-sample Hotelling's T-Squared [38]. AP

and PCKh are used as dependent variables, as AA and HOTA cannot be calculated for every tested algorithm. Specifically, we include AP for each separate subpose with ranged τ_{OKS}, and PCKh per keypoint for a total of 3 (subposes) + 17 (keypoints) = 20 parameters per sample. Each single-view video represents a sample for a total of 40 samples. We consider p-values of 0.05 or below to show statistical significance.

We repeat the same analysis to compare AlphaP^T and OpenPP^T on AA, where again the three ranged-τ_{OKS} subposes are used as separate input variables. Instead of repeating again with HOTA, we test on AA jointly with AP and/or PCKh.

QUALITATIVE ANALYSIS

To get insight into problems specific to our setting, results are manually evaluated. Specific example situations are selected by the authors to demonstrate strengths and weaknesses of each algorithm. Results are shown with detected poses, confidence scores and IDs. We show all detections regardless of their confidence.

3.2. RESULTS

This section shares the results obtained from the experiments described in section 3.1.3. Section 3.2.1 begins with an analysis of the dataset. Sections 3.2.2 to 3.2.5 report performance on various metrics. Statistical significance of the differences between algorithms is investigated in section 3.2.6. Finally, section 3.2.7 shows some qualitative examples.

3.2.1. DATASET COMPOSITION

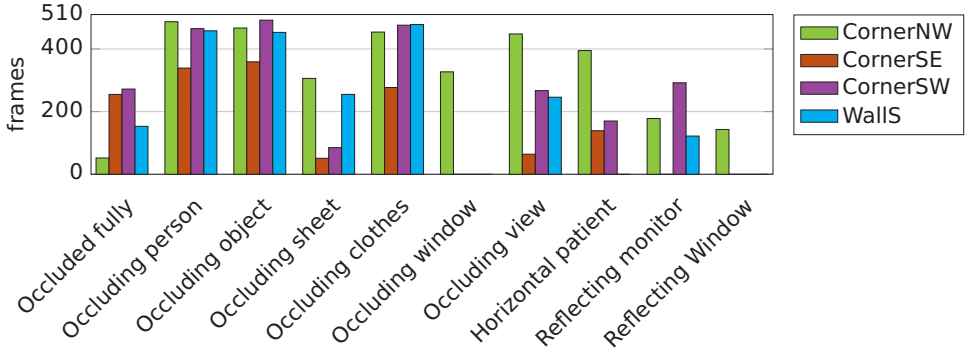


Figure 3.3.: Number of frames per situation from section 3.1.1 per viewpoint.

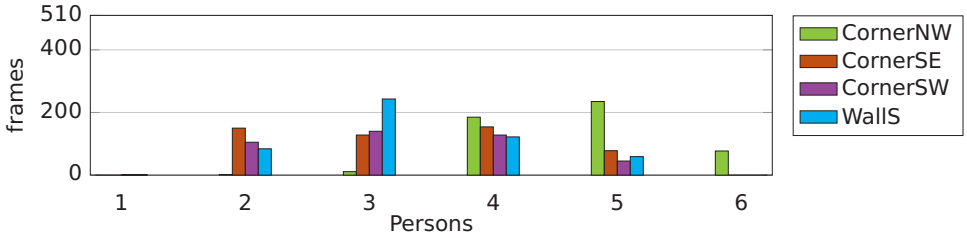


Figure 3.4.: Number of frames per person count per viewpoint.

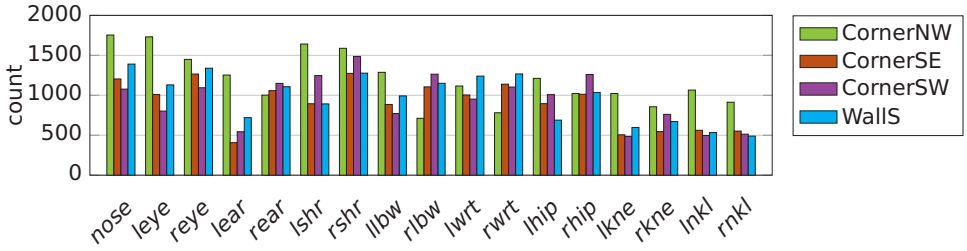


Figure 3.5.: Number of appearances per keypoint class from fig. 3.1 per viewpoint.

A description of the dataset is shown in figs. 3.3 to 3.5, as described in section 3.1.1. 1749 frames (85.7 % of the dataset) contain occlusion between persons, 1771 (86.8 %) occluding objects, and 1685 (82.6 %) occluding clothes. CornerNW, CornerSE, CornerSW and WallS saw 3257, 1484, 2519 and 2165 frame situations respectively, where counts exceed the dataset size due to single frames showing multiple situations. Most full occlusions and monitor reflections occur from CornerSW. Window occlusions- and reflections occur only from CornerNW. This viewpoint sees five persons on most frames and is the only viewpoint to ever see six. CornerSE usually sees four people and CornerSW and WallS three.

CornerNW, CornerSE, CornerSW and WallS respectively see a total of 20 404, 15 317, 16 012 and 16 518 keypoints. CornerNW sees the highest counts per subpose and class except for the *rear*, *rlbw* and *rhip*, which are seen more often from CornerSW, and the wrists which are seen more from WallS. CornerNW sees the highest mean count per class of 1200.2, paired with the highest standard deviation of 320.7. The lowest total and average counts are seen by CornerSE, although CornerSW sees lower counts per class more often.

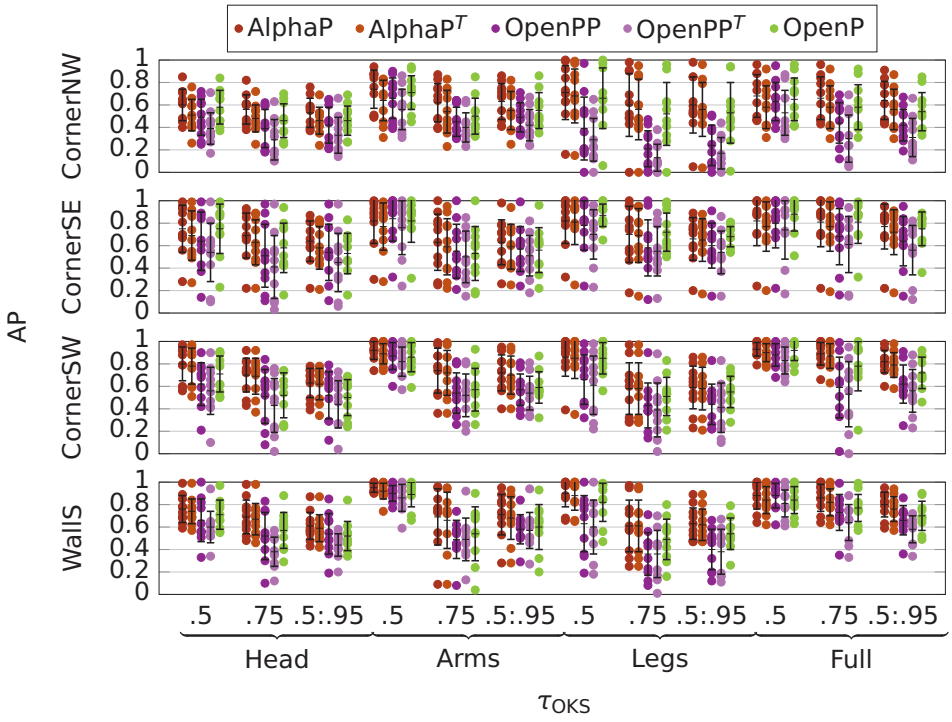


Figure 3.6.: Average Precision per individual video where error bars show two standard deviations around mean results.

3.2.2. AVERAGE PRECISION

AP can be seen in fig. 3.6. Here, AlphaP yields the highest Full-pose mean scores of up to 0.82. Non-tracking algorithms perform up to 11 percentage points (pp) better than their tracking counterparts. Arms yield the best scores of up to 0.72, and Head the worst of up to 0.64. AlphaP(^{T}) prefers the CornerSW viewpoint, OpenPP(^{T}) WallS, and OpenP CornerSE. CornerSW is most often preferred with up to 3 pp over the second-choice viewpoint per individual algorithm. On the Head and Arms, AlphaP(^{T}) shows scoring drops of up to 10 pp and 26 pp between low and high τ_{OKS}, which is 22 pp and 37 pp for other algorithms. On the Legs, OpenPP(^{T}) shows the lowest scoring drop of up to 27 pp which is 34 pp for others. Results on the Arms show standard deviations of up to 20 pp. For the Head and Legs this is 26 pp and 27 pp respectively. OpenP shows standard deviations of up to 17 pp, AlphaP(^{T}) of 20 pp and OpenPP(^{T}) of 22 pp.

From CornerSE, one outlier procedure performs worse than the rest for all algorithms. Here the cardiologist and patient are mostly occluded by the monitor. Their few visible keypoints were not detected, or merged into a single pose.

3.2.3. HEAD-GUIDED PERCENTAGE OF CORRECT KEYPOINTS

PCKh in fig. 3.7 shows that most keypoints prefer OpenPP or OpenPP^T except the *nose*, which prefers AlphaP instead. All algorithms prefer WallS most often, followed by CornerSE for AlphaP and OpenPP, and CornerSW for AlphaP^T and OpenP. AlphaP, OpenPP^T and OpenP prefer CornerNW least often, which is CornerSE for AlphaP^T and CornerSW for OpenPP. With scores of up to 0.57 and 0.87 the Legs and Head score the lowest and highest respectively. The *hip*s are detected worst with a maximum score of 0.23. All subposes prefer WallS. All algorithms had the highest standard deviation on CornerSW. The lowest standard deviations for the Head are achieved on CornerSE and WallS, for the Arms on WallS, and for the Legs on CornerNW.

The outlier procedure from CornerSE at the end of section 3.2.2 shows the same poor performance on PCKh. From CornerSW, we see another procedure scoring below the others. This video shows two people standing close together, dressed in loose medical aprons and facing away from the camera whilst the instrument table occludes their legs.

3.2.4. ASSOCIATION ACCURACY

Looking at tracking, fig. 3.8 shows that AlphaP^T outperforms OpenPP^T on mean Full-pose AA from all viewpoints except WallS. Arms are tracked best in most situations, and Legs the worst. AlphaP^T shows little mean Full-pose scoring drop of up to 2 pp between low and high τ_{OKS}, which

Figure 3.7.: Head-guided Percentage of Correct Keypoints over the entire dataset where error bars show two standard deviations around mean results.

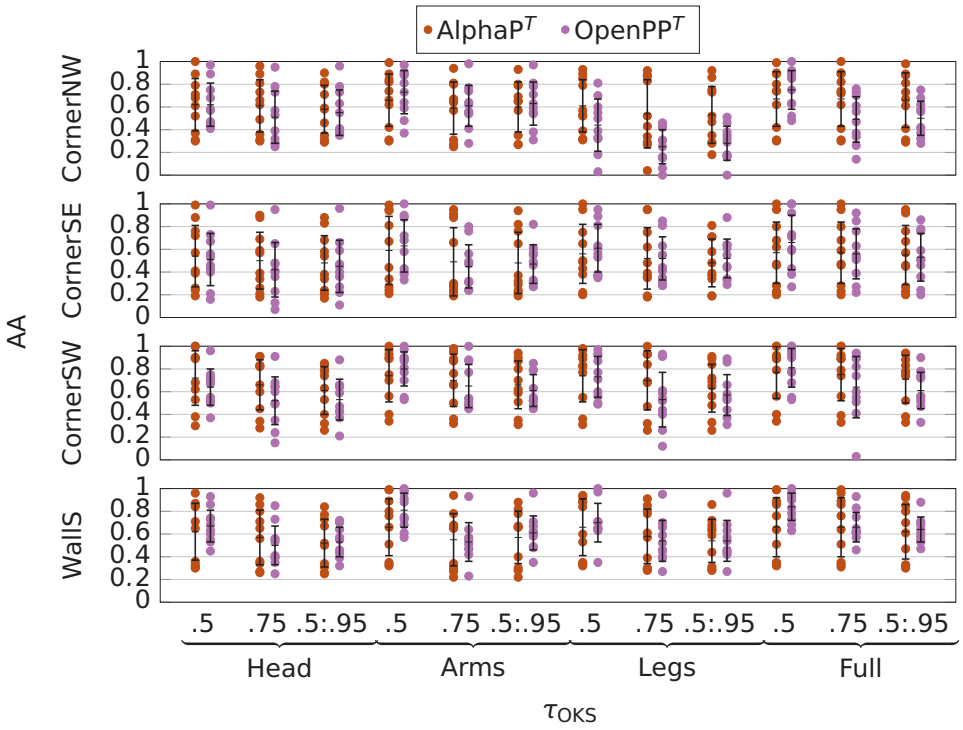


Figure 3.8.: Association Accuracy per viewpoint and subpose over the entire dataset where error bars show two standard deviations around mean results.

is 26 pp for OpenPP^T. On AA this drop is larger for OpenPP^T than for AlphaP^T for all subposes and viewpoints. However, OpenPP^T yields lower standard deviation than AlphaP^T for all subposes and the Full pose from all viewpoints.

3.2.5. HIGHER-ORDER TRACKING ACCURACY

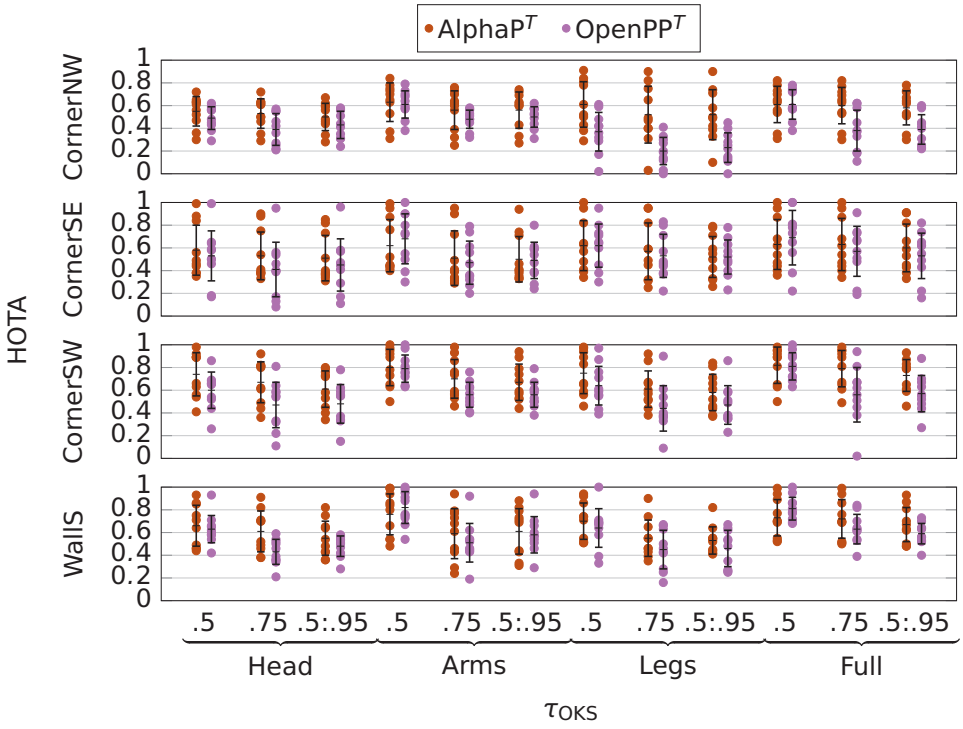


Figure 3.9.: Higher-Order Tracking Accuracy per viewpoint and subpose over the entire dataset where error bars show two standard deviations around mean results.

The integration of tracking and detection metrics with HOTA in fig. 3.9 sees AlphaP^T outperform OpenPP^T everywhere except with low τ_{OKS} for some subposes and viewpoints. OpenPP^T still yields lower standard deviations except for the Head from CornerSE and CornerSW, Legs from CornerSW and WallS, and Full pose from CornerSW. The highest achieved Full-pose mean scores are 0.73 for AlphaP^T and 0.59 for OpenPP^T. Arms and CornerSW are preferred in most situations.

Table 3.2.: p-value per algorithm pair from Hotelling's T-Squared with AP and PCKh as parameters.

	AlphaP^T	OpenPP	OpenPP^T	OpenP
AlphaP	0.9999	<0.0001	<0.0001	<0.0001
AlphaP^T		<0.0001	<0.0001	<0.0001
OpenPP			0.9103	0.0003
OpenPP^T				<0.0001

3.2.6. HOTELLING'S T-SQUARED

Calculated p-values are shown in table 3.2 per pair of algorithms. The only non-significant differences occur when comparing tracking- and non-tracking versions of the same algorithm, in which case p-values approach 1.

When excluding AP from the test, conclusions remain the same except for a now statistically insignificant p-value between OpenPP and OpenP. Excluding PCKh instead gives insignificance between AlphaP^T and OpenP.

Testing on AA yields an insignificant difference between AlphaP^T and OpenPP^T. Adding AP and/or PCKh lowers the p-value back below our significance threshold.

3.2.7. QUALITATIVE RESULTS

Example detections are demonstrated in fig. 3.10. In the first column people are standing close together. OpenPP^T merges the patient and cardiologist with 0.87 confidence. AlphaP(^{T}) mistakes the patient as part of the cardiologist at 0.72. The cardiologist and assistant who stands close are never merged. Only OpenPP and OpenP see the patient and merge no-one. OpenP detects most correct poses, but with the lowest confidence. All models except OpenPP^T are least confident about the cardiologist, who faces away from the camera.

The second column shows the cardiologist putting on an apron. AlphaP places a full pose with confidence 0.31 where only his Head is visible, and AlphaP^T sees nothing. OpenPP(^{T}) correctly detects the *shr*s and *hip*s at 0.82, although *hip* placements seem off. OpenPP^T additionally sees a lower arm in the sleeve. Only OpenP detects the *nkl*s. All algorithms detect the lab assistant where AlphaP, AlphaP^T and OpenP place the occluded *lnkl* wrongly with 0.84, 0.84 and 0.69 confidence.

In the third column the monitor reflects a lab assistant. All algorithms detect the reflection, where OpenPP is most confident at 0.83 and OpenP the least at 0.63. AlphaP, AlphaP^T and OpenP hallucinate two *kne*s and/or *nkl*s. These models detect the full cardiologist at 0.74, 0.74

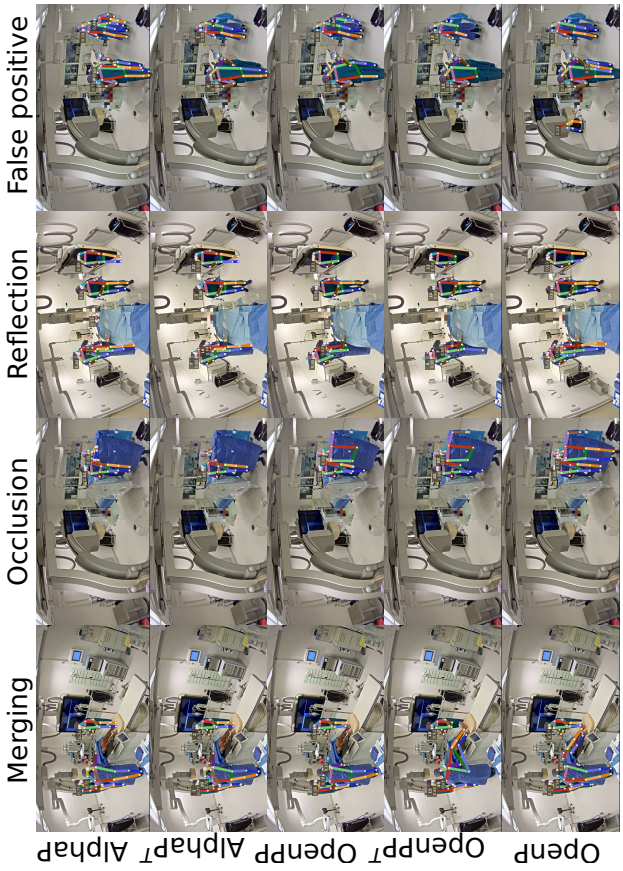


Figure 3.10.: Qualitative detections with confidences and tracking IDs.

and 0.60 confidence, where OpenPP(^{T}) detects all but his legs at 0.87 and 0.69. All models see the occluded assistant, where OpenPP(^{T}) is most confident at 0.95 and AlphaP(^{T}) the least at 0.62. Similarly to column two, AlphaP incorrectly detects a full pose around the head of the patient with 0.31 confidence. OpenPP correctly detects only their Head keypoints at 0.90.

The last column shows the instrument table with a sheet resembling clothing. AlphaP and OpenPP detect a pose here at 0.44 and 0.64 confidence. The occluding assistants are fully detected by AlphaP at 0.72 and 0.66 and merged by OpenPP at 0.77. AlphaP^T and OpenPP^T only detect the closest assistant at 0.72 and 0.90. OpenP detects both assistants partially at 0.73 and 0.57. It also sees Legs in the background bin with 0.22 certainty.

Figure 3.11 shows tracking results from AlphaP^T and OpenPP^T. The first and third row show the cardiologist and assistant preparing, with a patient on the table. AlphaP^T detects the partial cardiologist at the bottom in 3 frames, which is only 1 for OpenPP^T. OpenPP^T detects the patient more consistently and with higher confidence. AlphaP^T moves ID 1 from the assistant to the cardiologist. OpenPP^T yields no such identity swaps.

In the remaining rows the cardiologist exits and re-enters the room, with the patient waiting in the hallway. After their return, OpenPP^T assigns the cardiologist a new ID whereas AlphaP^T recognises them from before. The same happens for the patient after the cardiologist passes them in front. When watching frame by frame, AlphaP^T shows many identity swaps even with just one person visible.

3.3. DISCUSSION

In this paper we introduced a dataset with footage from real CAG procedures in the Cath Lab, and provided benchmark results of several pose estimation- and tracking algorithms. Quantitative metrics were evaluated on our annotated dataset per subpose and viewpoint, and qualitative observations were shown.

We observe that AlphaP(^{T}) produces the best AP, whereas OpenPP(^{T}) performs better on PCKh. As AP is calculated on (sub)poses and PCKh per keypoint, this suggests that OpenPP(^{T}) places keypoints more accurately and AlphaP(^{T}) connects them into poses better. This is in line with the top-down approach of AlphaP, which applies local restrictions on matchable keypoint pairs. Another explanation is that AlphaP(^{T}) could score poses more accurately in the Cath Lab, as AP considers confidence score. This could explain why metrics on AlphaP(^{T}) do not drop much between low and high τ_{OKS}, as accurate scoring might compensate cases of poor localization. Measured AP scores are higher in the Cath Lab than those reported on the MVOR dataset [24], although different

Figure 3.11.: Qualitative tracklets with confidences and tracking IDs.

recording methods render these results not directly comparable.

AP and PCKh show differences per viewpoint and keypoint. Legs especially are subject to occlusion in clinical settings. Head keypoints are hidden behind masks and hairnets. Different viewpoints see different levels of such occlusion, and therefore show varying results. AlphaP(^{T}) tends to detect the Head better, possibly by imposing a prior through object detection. Qualitative results show a drawback of this approach, where priors encourage placement of full poses on partially visible humans or inanimate objects. We do observe that these incorrect keypoints receive low confidences, which is in line with the theory that AlphaP yields more accurate scores. Hence, in practice this drawback poses little issue if confidence is considered appropriately.

On AA, AlphaP^T scores better than OpenPP^T for our dataset. OpenPP^T tends to miss people in our setting or merge them; possibly due to the temporal limbs providing more matching paths to do so in close proximity. AlphaP^T still scored poorly with large variability between workflow phases. This could be due to its use of visual clues, which in combination with indistinguishable sterile clothing may have caused the many identity swaps. These issues could explain why the tracking models were outperformed by their non-tracking counterparts on AP, although this difference was only small and proved insignificant.

HOTA eases comparison by integrating detection and tracking performance. Here, AlphaP^T slightly outperforms OpenPP^T. When looking purely at this combined metric, using AlphaP^T from the CornerSW viewpoint seems to perform best in the Cath Lab.

Procedures are carried out with Arms, and Head orientation indicates where one is focussing. In our setting, Legs serve only to reposition oneself; something that can be inferred from other keypoints. Therefore for workflow analysis, Arms movement is probably the most descriptive followed by the Head and then Legs. Hence, we should prioritise subpose detections in that order.

Monitor reflections and occlusion pose problems for pose detection- and tracking in the Cath Lab. Reflections are a problem because the tested detectors are not trained to distinguish them from real human beings [39]. For workflow purposes the activity of persons is of interest, and their reflections serve only as noise. Occlusion renders persons invisible from individual views, causing False Negatives, or causing detected body joints to be connected incorrectly. Especially during tracking this presents an issue, as reidentification is difficult after losing- or wrongfully detecting a person. Tracking algorithms solve this problem through visual reidentification, but that does not work in the Cath Lab where everyone is dressed similarly.

CornerSW and WallS yield the best results in most situations. Although monitor reflections plague both, their limited occlusion and view of only the Cath Lab interior simplify the problem. CornerSE sees no reflections,

but suffers from occlusions in the patient area by the monitor and operating table. CornerNW sees occlusion from the radiation shield and C-arm, reflections in the control room window and monitor, and people in the control room whose movements can be assumed to provide no relevant workflow information. The cardiologist facing away from CornerSW makes the use of this view for workflow analysis questionable. WallS, with its clear yet narrow view on the operating table surroundings, is an intuitive choice for workflow analysis during procedures. Before and after procedures, CornerSE provides a clearer view around the room entrances.

Human movement is descriptive of personnel activities [10–12], making reliable tracking important for workflow analysis. Unfortunately, no tested model yielded good tracking results on our dataset. AlphaP^T produced multiple identity swaps per minute and OpenPP^T merged or missed people.

Our model selection was limited with only one top-down algorithm and the dataset was relatively small. We did not annotate occluded keypoints which may have unfairly increased scores for more occluded viewpoints.

3.3.1. FUTURE RESEARCH

In following studies, more estimators [16, 17, 20] could be tested. Tracking should be done with a separate algorithm for better tracking performance. To overcome the visual differences between clothing in the Cath Lab and in general datasets like COCO, domain adaptation- or generalization methods could be explored [40]. With enough annotations, models could be re-trained for the Cath Lab or specific subposes. Interesting would be to annotate and detect keypoints in the C-arm, table, or lead screen. Expanding from single-view to multiple-view or 3D pose detection could help mitigate occlusion, as explored for the OR in [25–27]. It can be investigated how yielded poses can be used for automated recognition of e.g. personnel activities, workflow phases, or radiation exposure.

The insights from this work can aid the design of new computer vision setups in the Cath Lab or OR. For instance, cameras are best placed in a position which provides a clear view on personnel from the front, excluding reflective monitors or windows. As occlusion can rarely be avoided, a clear view of the Arms should be prioritised. When exploring pose detection, an algorithm can be chosen based on discussed trade-offs. In the design of a new pose detector one could focus on robustness against Cath Lab-specific occlusion, or distinguishing between real poses and reflections. When tracking, it is probably best not to use visual features due to similarities in appearance from sterile clothing.

The study shows that, considering confidence scoring and keypoint

matching, AlphaP is the best-suited tested model in the Cath Lab. When only keypoint locations are sought, OpenPP could be a better choice. Due to identity swaps and pose merging, no tested tracker seems sufficient for use in workflow analysis. A new tracker should be developed based on the shortcomings highlighted in this paper. It should address the visual complexity of the Cath Lab specifically.

3.4. CONCLUSIONS

We annotated poses and identities in 2040 frames from ten CAG procedures. Detection- and tracking metrics AP, PCKh, AA and HOTA were calculated for the models from table 3.1. Models showed significant performance differences, except when comparing different models of the same algorithm. The WallS and CornerSW viewpoints from fig. 3.2 and the Arms keypoints were scored highest upon. The room coverage and decent results of CornerSE make this view a suitable alternative for workflow analysis, although its results vary with monitor positioning. OpenPP produced the most accurate keypoint locations in the Cath Lab. AlphaP(^{T}) yielded the best confidence scores, keypoint matching, and tracking results.

REFERENCES

[1] R. M. Butler, E. Frassini, T. S. Vijfvinkel, S. van Riel, C. Bachvarov, J. Constandse, M. van der Elst, J. J. van den Dobbelsteen, and B. H. W. Hendriks. “Benchmarking 2D Human Pose Estimators and Trackers for Workflow Analysis in the Cardiac Catheterization Laboratory”. In: *Med. Eng. Phys.* 136 (Feb. 2025), p. 104289. doi: 10.1016/j.medengphy.2025.104289.

[2] K. N. Timoh, A. Huaulme, K. Cleary, M. A. Zaheer, V. Lavoué, D. Donoho, and P. Jannin. “A Systematic Review of Annotation for Surgical Process Model Analysis in Minimally Invasive Surgery based on Video”. In: *Surg. Endosc.* 37 (May 2023), pp. 4298–4314. doi: 10.1007/s00464-023-10041-w.

[3] A. M. Schouten, S. M. Flipse, K. E. van Nieuwenhuizen, F. W. Jansen, A. C. van der Eijk, and J. J. van den Dobbelsteen. “Operating Room Performance Optimization Metrics: a Systematic Review”. In: *J. Med. Syst.* 47.1 (Dec. 2023), p. 19. doi: 10.1007/s10916-023-01912-9.

[4] C. R. Garrow, K.-F. Kowalewski, L. Li, M. Wagner, M. W. Schmidt, S. Engelhardt, D. A. Hashimoto, H. G. Kenngott, S. Bodenstedt, S. Speidel, B. P. Müller-Stich, and F. Nickel. “Machine Learning for Surgical Phase Recognition: A Systematic Review”. In: *Annal. Surg.* 273.4 (Apr. 2021), pp. 684–693. doi: 10.1097/SLA.0000000000004425.

[5] F. Lalys and P. Jannin. “Surgical Process Modelling: a Review”. In: *Int. J. Comput. Assist. Radiol. Surg.* 9 (May 2014), pp. 495–511. doi: 10.1007/s11548-013-0940-5.

[6] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, M. Hashizume, D. Katic, H. Kenngott, M. Kranzfelder, A. Malpani, K. März, T. Neumuth, N. Padoy, C. Pugh, N. Schoch, D. Stoyanov, R. Taylor, M. Wagner, G. D. Hager, and P. Jannin. “Surgical Data Science for Next-generation Interventions”. In: *Nat. Biomed. Eng.* 1 (Sept. 2017), pp. 691–696. doi: 10.1038/s41551-017-0132-7.

[7] E. Bkheet, A.-L. D'Angelo, A. Goldbraikh, and S. Laufer. "Using Hand Pose Estimation to Automate Open Surgery Training Feedback". In: *Int. J. Comput. Assist. Radiol. Surg.* 18 (May 2023), pp. 1279–1285. doi: 10.1007/s11548-023-02947-6.

[8] I. Aksamentov, A. P. Twinanda, D. Mutter, J. Marescaux, and N. Padoy. "Deep Neural Networks Predict Remaining Surgery Duration from Cholecystectomy Videos". In: *Med. Image Comput. Comput.-Assist. Interv.* Springer, Sept. 2017, pp. 586–593. doi: 10.1007/978-3-319-66185-8_66.

[9] M. Berlet, T. Vogel, D. Ostler, T. Czempiel, M. Kähler, S. Brunner, H. Feussner, D. Wilhelm, and M. Kranzfelder. "Surgical Reporting for Laparoscopic Cholecystectomy based on Phase Annotation by a Convolutional Neural Network (CNN) and the Phenomenon of Phase Flickering: a Proof of Concept". In: *Int. J. Comput. Assist. Radiol. Surg.* 17 (Nov. 2022), pp. 1991–1999. doi: 10.1007/s11548-022-02680-6.

[10] G. Saleem, U. I. Bajwa, and R. H. Raza. "Toward Human Activity Recognition: a Survey". In: *Neural Comput. Appl.* 35 (Feb. 2023), pp. 4145–4182. doi: 10.1007/s00521-022-07937-4.

[11] H.-C. Nguyen, T.-H. Nguyen, and V.-H. Scherer Rafałand Le. "Deep Learning for Human Activity Recognition on 3D Human Skeleton: Survey and Comparative Study". In: *Sens.* 23.11 (May 2023), p. 5121. doi: 10.3390/s23115121.

[12] C. Wang and J. Yan. "A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition". In: *IEEE Access* 11 (June 2023), pp. 53880–53898. doi: 10.1109/ACCESS.2023.3282311.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. "Microsoft COCO: Common Objects in Context". In: *Eur. Conf. Comput. Vis.* Springer, Sept. 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[14] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye. "Object Detection in 20 Years: A Survey". In: *Proc. IEEE* 111.3 (Jan. 2023), pp. 257–276. doi: 10.1109/JPROC.2023.3238524.

[15] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu. "AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 45.6 (June 2023), pp. 7157–7173. doi: 10.1109/TPAMI.2022.3222784.

[16] K. Sun, B. Xiao, D. Liu, and J. Wang. "Deep High-Resolution Representation Learning for Human Pose Estimation". In: *Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2019, pp. 5686–5696. doi: 10.1109/CVPR.2019.00584.

[17] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele. “ArtTrack: Articulated Multi-Person Tracking in the Wild”. In: *IEEE Conf. Comput. Vis. Pattern. Recognit.* IEEE, July 2017, pp. 1293–1301. doi: 10.1109/CVPR.2017.142.

[18] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 43.1 (Jan. 2021), pp. 172–186. doi: 10.1109/TPAMI.2019.2929257.

[19] S. Kreiss, L. Bertoni, and A. Alahi. “OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association”. In: *IEEE Trans. Intell. Transp. Syst.* 23.8 (Aug. 2022), pp. 13498–13511. doi: 10.1109/TITS.2021.3124981.

[20] Z. Geng, K. Sun, B. Xiao, Z. Zhang, and J. Wang. “Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression”. In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2021, pp. 14671–14681. doi: 10.1109/CVPR46437.2021.01444.

[21] Y. Zhang, P. Sun, y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang. “ByteTrack: Multi-object Tracking by Associating Every Detection Box”. In: *Eur. Conf. Comput. Vis.* Springer, Oct. 2022, pp. 1–21. doi: 10.1007/978-3-031-20047-2_1.

[22] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. “Multiple Object Tracking Using K-Shortest Paths Optimization”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 33.9 (Sept. 2011), pp. 1806–1819. doi: 10.1109/TPAMI.2011.21.

[23] L. Bastian, T. D. Wang, T. Czempiel, B. Busam, and N. Navab. “DisguisOR: Holistic Face Anonymization for the Operating Room”. In: *Int. J. Comput. Assist. Radiol. Surg.* 18 (July 2023), pp. 1209–1215. doi: 10.1007/s11548-023-02939-6.

[24] V. Srivastav, T. Issenhuth, K. Abdolrahim, M. de Mathelin, A. Gangi, and N. Padoy. “MVOR: A Multi-View RGB-D Operating Room Dataset for 2D and 3D Human Pose Estimation”. In: *Conf. Med. Image Comput. Comput. Assist. Interv.* MICCAI, 2018.

[25] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin, and N. Padoy. “Articulated Clinician Detection using 3D Pictorial Structures on RGB-D Data”. In: *Med. Image Anal.* 35 (Jan. 2017), pp. 215–224. doi: 10.1016/j.media.2016.07.001.

[26] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin, and N. Padoy. “A Multi-View RGB-D Approach for Human Pose Estimation in Operating Rooms”. In: *IEEE Winter Conf. Appl. Comput. Vis.* IEEE, Mar. 2017, pp. 363–372. doi: 10.1109/WACV.2017.47.

[27] A. Kadkhodamohammadi and N. Padoy. “A Generalizable Approach for Multi-View 3D Human Pose Regression”. In: *Mach. Vis. Appl.* 32 (Oct. 2021), p. 6. doi: 10.1007/s00138-020-01120-2.

[28] Z. Wang, R. Butler, J. J. van den Dobbelsteen, B. H. W. Hendriks, M. van der Elst, and J. Dauwels. “Towards Robust Object Detection in Unseen Catheterization Laboratories”. In: *IEEE Int. Workshop Med. Meas. Appl.* IEEE, June 2024. doi: 10.1109/MeMeA60663.2024.10596906.

[29] CVAT.ai Corporation. *Computer Vision Annotation Tool (CVAT)*. url: https://www.cvat.ai.

[30] J. Redmon and A. Farhadi. *YOLOv3: An Incremental Improvement*. Apr. 2018. doi: 10.48550/arXiv.1804.02767. arXiv: 1804.02767v1 [cs.CV].

[31] K. He, X. Zhang, S. Ren, and J. Sun. “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 37.9 (Sept. 2015), pp. 1904–1916. doi: 10.1109/TPAMI.2015.2389824.

[32] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *Proc. 29th IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[33] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. “ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design”. In: *Proc. 15th Eur. Conf. Comput. Vis.* Springer, Sept. 2018, pp. 122–138. doi: 10.1007/978-3-030-01264-9_8.

[34] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. da Silva. “A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit”. In: *Electron.* 10.3 (Jan. 2021), p. 279. doi: 10.3390/electronics10030279.

[35] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. “2D Human Pose Estimation: New Benchmark and State of the Art Analysis”. In: *IEEE Conf. Comput. Vis. Pattern. Recognit.* IEEE, June 2014, pp. 3686–3693. doi: 10.1109/CVPR.2014.471.

[36] H. W. Kuhn. “Variants of the Hungarian Method for Assignment Problems”. In: *Nav. Res. Logist. Q.* 03.04 (Dec. 1956), pp. 253–258. doi: 10.1002/nav.3800030404.

[37] J. Luiten, A. Osšep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe. “HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking”. In: *Int. J. Comput. Vis.* 129.2 (Feb. 2021), pp. 548–578. doi: 10.1007/s11263-020-01375-2.

[38] H. Hotelling. “The Generalization of Student’s Ratio”. In: *Ann. Math. Stat.* 2.3 (Aug. 1931), pp. 360–378.

[39] D. Park and Y.-H. Park. “Identifying Reflected Images from Object Detector in Indoor Environment utilizing Depth Information”. In: *IEEE Robot. Autom. Lett.* 6.2 (Apr. 2021), pp. 635–642. doi: 10.1109/LRA.2020.3047796.

[40] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, and W. Lu. “Generalization to Unseen Domains: A Survey on Domain Generalization”. In: *IEEE Trans. Knowl. Data Eng.* 35.8 (Aug. 2023), pp. 8052–8072. doi: 10.1109/TKDE.2022.3178128.


4

POSEBYTE: ROBUST 2D HUMAN POSE TRACKING

Workflow insights can enable safety- and efficiency improvements in the Cardiac Catheterisation Laboratory (Cath Lab). Human pose tracklets from video footage can provide a source of workflow information. However, occlusions and visual similarity between personnel make the Cath Lab a challenging environment for the reidentification of individuals. We propose a human pose tracker that addresses these problems specifically, and test it on recordings of real coronary angiograms. This tracker uses no visual information for reidentification, and instead employs object keypoint similarity between detections and predictions from a third-order motion model. Algorithm performance is measured on Cath Lab footage using Higher-Order Tracking Accuracy (HOTA). To evaluate its stability during procedures, this is done separately for five different surgical steps of the procedure. We achieve up to 0.71 HOTA where tested state-of-the-art pose trackers score up to 0.65 on the used dataset. We observe that the pose tracker HOTA performance varies with up to 10 percentage point (pp) between workflow phases, where tested state-of-the-art trackers show differences of up to 23 pp. In addition, the tracker achieves up to 22.5 frames per second, which is 9 frames per second faster than the current state-of-the-art on our setup in the Cath Lab. The fast and consistent short-term performance of the provided algorithm makes it suitable for use in workflow analysis in the Cath Lab and opens the door to real-time use-cases. Our code is publicly available at https://github.com/RM-8vt13r/PoseBYTE.

Parts of this chapter were published in Med. Eng. Phys. **135** (2025), © Elsevier [1].

The emerging field of workflow analysis promises tools for the analysis and improvent of surgical procedures [2–4]. Insights into workflow could be used to improve e.g. procedure efficiency and safety through personnel training. We investigate a tool for workflow analysis in the Cardiac Catheterisation Laboratory (Cath Lab): a specialised Operating Room (OR) for minimally invasive cardiovascular procedures. The Cath Lab is equipped for its specialised purpose with a fixed X-Ray imaging system containing a 'C-Arm' mount, a monitor, and a radiation shield.

One diagnostic procedure carried out in the Cath Lab is the coronary angiogram (CAG) [5]. During a CAG, cardiovascular access is established through the wrist or groin area using a catheter. A contrast fluid is administered directly into the coronary artery to detect anomalies on a captured X-Ray image. Reference [6] provides a description of the CAG in terms of consequtive workflow steps. Its well-defined nature makes the CAG suitable for explorative workflow study.

Manual workflow recognition is labour-intensive. In contrast, computer-assisted automation [7–9] is cost-effective, scalable, and enables real-time use-cases and assistance [10]. Multi-object keypoint detection can serve as a stepping stone to activity recognition [11–13]. 2D keypoint detectors—or pose estimators—localise predefined objects and their keypoints in continuous image pixel (px) space. They quantify detection confidence with a score per detected keypoint.

Multi-Object Tracking builds on detection by assigning the same identifier (ID) to the same object in different video frames. A tracker outputs a set of tracklets, each of which contains the per-frame detections of a unique object.

Many human pose tracking algorithms exist [14–17], which were benchmarked in general environments [18]. Several existing human pose trackers wrongfully swap identities or merge pose in the Cath Lab, as personnel occlusion and visual similarity are common.

In this paper we adapt BYTE [19]—a state-of-the-art bounding box tracker—for pose tracking in the Cath Lab. BYTE reidentifies objects or persons by comparing bounding box detections on subsequent frames. Persons pass each other regularly in the Cath Lab, after which their bounding boxes will be hard to distinguish from geometry alone. The visual features that BYTE uses to mitigate this problem are less effective here than in the general case, because everyone is dressed very similarly. Poses provide more geometric information that can be used for reidentification than a bounding box, by specifying keypoint coordinates. Therefore, we replace the use of bounding boxes in BYTE by human poses such that, after or during occlusion by a person or object, a person can be reidentified by posture. Additionally we extend the constant-velocity motion model that BYTE uses with acceleration and jerk to model more complex movement. These changes mitigate occlusion-induced problems like identity swaps or lost tracklets. As

visual similarity between personnel can cause identity swaps, the tracker uses no image data. In the remaining text, we refer to the proposed method as 'PoseBYTE', indicating its utilisation of human pose data rather than bounding boxes for reidentification.

CAG workflow phases [6] differ in terms of appearance and movement. For instance, whilst the patient walks to the operating table there is a lot of movement and occlusion from ongoing preparations. During the intervention there are fewer people and less walking, and more subtle hand- and head motion. The lights being switched on or off during different phases causes visual differences. For accurate workflow analysis from poses, it is important that a pose tracker works throughout a procedure. Therefore, we test PoseBYTE separately during five different workflow phases.

Annotated video data are necessary to test pose trackers. We use 30-second video sequences of five CAG workflow phases from the Cath Lab of the Reinier de Graaf Gasthuis hospital, Delft, NL, all filmed from four viewpoints. Ground-truth human pose tracklets were annotated in the footage to evaluate metrics.

Section 4.1 starts with a description of our dataset, algorithm and experiments. section 4.2 lists results and discusses those that stand out. These are further interpreted in section 4.3. Finally, section 4.4 gives a summary.

4.1. TRACKING ALGORITHM DESIGN

4.1.1. DATASET

The recording of CAG procedures in the Reinier de Graaf Gasthuis hospital, Delft, NL was approved by the Medical Ethics Committee Leiden The Hague Delft (protocol number Z19.057, 30-10-2019) and the hospital board. Informed consent was collected from all filmed patients and staff. Procedures were recorded in the hospital Cath Lab from four different viewpoints (Axis M1125) in a resolution of $1920\,\text{px} \times 1080\,\text{px}$ and framerate of 25 frames per second (fps). A cardiologist, scrub nurse, up to two lab assistants, and the patient were present during each procedure.

We annotate poses in ten procedure recordings, each performed by a different medical team for variability. 51 frames were sampled uniformly over 30 seconds per procedure, from each of the four synchronised viewpoints. This gives a total of $10 \times 51 \times 4 = 2040$ annotated frames. The video sequences were hand-selected to show five different workflow phases, each taken from two different procedures:

- the patient entering and lying down,
- realisation of endovascular access,

- Use of ultrasound to detect the radial artery for endovascular access,
- X-Ray imaging, and
- closure of the entrywound.

The annotations were made in Computer Vision Annotation Tool (CVAT) [20][1] by two authors with an engineering background. Annotation quality was checked by another author who has been a practicing interventional cardiologist for over 13 years. Occluded persons and reflections in the monitor or windows were not labelled.

4.1.2. POSE DETECTION

2D human poses are detected per frame with a keypoint detector and serve as input to PoseBYTE. Section 4.1.4 discusses the tested detectors.

4.1.3. POSEBYTE

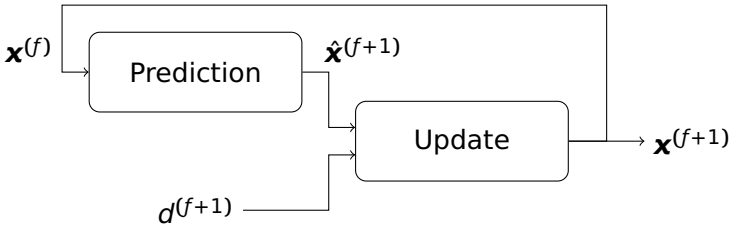


Figure 4.1.: Schematic of a Kalman filter [21]. A state $\boldsymbol{x}^{(f)}$ is kept internally. Given a noisy measurement $d^{(f+1)}$ and a model prediction $\hat{\boldsymbol{x}}^{(f+1)}$ based on $\hat{\boldsymbol{x}}^{(f)}$, a new state $\boldsymbol{x}^{(f+1)}$ is produced. $\boldsymbol{x}^{(f+1)}$ contains a denoised measurement and estimated hidden variables that are used in the prediction model.

BYTE [19] is a state-of-the-art tracking algorithm that reidentifies bounding boxes produced by an object detector [22] between frames. It keeps a set of tracklets, each of which keeps a state column vector with its position $\boldsymbol{p}^{(f)}$ [px] and velocity $\boldsymbol{v}^{(f)}$ [px f^{-1}] (pixels per frame)

$$\boldsymbol{x}^{(f)} = \begin{bmatrix} \boldsymbol{p}^{(f)} \\ \boldsymbol{v}^{(f)} \end{bmatrix} \tag{4.1}$$

per frame f. Here, the unit f is the time duration of a single frame. On each frame, a Kalman filter [21] produces a model-based prediction

[1] Code available: https://github.com/cvat-ai/cvat

$\hat{\boldsymbol{p}}^{(f+1)} = \boldsymbol{p}^{(f)} + \boldsymbol{v}^{(f)}$ as visualised in fig. 4.1. This is called the prediction step.

Bounding box detection confidences are classified as high or low with a threshold γ_{high}. A similarity metric, e.g., Intersection over Union (IoU) [23] or a visual reidentification feature, measures resemblance between high-confidence detections and tracklet predictions $\hat{\boldsymbol{p}}^{(f+1)}$. Similarity scores above threshold σ_{high} are used in the Hungarian algorithm [24] to match detections to predictions. The Kalman filter provides an updated state $\boldsymbol{x}^{(f+1)}$ from each match, shown in the right part of fig. 4.1. This is called the update step.

Low-confidence detections are matched to remaining tracklets with a similarity metric that does not rely on visual information. Here, another similarity score threshold $\sigma_{\text{low}} < \sigma_{\text{high}}$ is applied. Unmatched tracklets are labelled as 'lost', their new state being predicted on each frame until i) they can be matched to a detection and the tracklet continues, or ii) f_{mem} frames have passed and the tracklet ends. Remaining high-confidence boxes seed new tracklets, which are confirmed on the next frame with a similarity threshold $\sigma_{\text{new}} < \sigma_{\text{high}}$ before proceeding as usual.

4

POSE TRACKING

We adapt the Kalman filter to store coordinates and velocities per keypoint rather than per object. Position and velocity in eq. (4.1) become

$$\boldsymbol{p}^{(f)} = \begin{bmatrix} x_1^{(f)} \\ y_1^{(f)} \\ \vdots \\ x_{|\mathcal{K}|}^{(f)} \\ y_{|\mathcal{K}|}^{(f)} \end{bmatrix}, \boldsymbol{v}^{(f)} = \begin{bmatrix} vx_1^{(f)} \\ vy_1^{(f)} \\ \vdots \\ vx_{|\mathcal{K}|}^{(f)} \\ vy_{|\mathcal{K}|}^{(f)} \end{bmatrix}, \tag{4.2}$$

where $\mathcal{K}$ and $|\square|$ denote the set of keypoint classes and set cardinality operator, and $x_k^{(f)}$, $y_k^{(f)}$, $vx_k^{(f)}$, $vy_k^{(f)}$ are horizontal (x) and vertical (y) position and speed of keypoint $k \in \mathcal{K}$ on frame f.

If and only if i) a tracklet and a pose detection are matched, and ii) a keypoint in the pose has a confidence below γ_{kp}, we exclude this keypoint from the update step. This is done by leaving out the rows corresponding to this keypoint in the Kalman filter observation matrix and observation vector during the update. Thus, in this case, the state of this keypoint on the next frame is purely its model-based prediction. If a keypoint has a confidence below γ_{kp} when starting a new tracklet, we apply a large 10 000 px observation uncertainty to it instead as we need to initialise a full initial state. When a tracklet is not matched or no keypoints remain after thresholding, the tracklet is lost but can be found

again as described in section 4.1.3. Whilst a tracklet is lost, predicted keypoints act for future matching only and are not saved as part of the tracklet. Tracklet keypoint coordinates are taken from the Kalman filter update step, and confidences copied from the detector.

We use the mean of all nonzero keypoint confidences as pose score, and the tightest-fit bounding box as approximate segmentation area. Object Keypoint Similarity (OKS) [25] is used as Similarity score, in which calculation the Kalman filter prediction is treated as ground truth. We do not use any visual clues, as similarities between personnel can make this an unreliable feature for pose tracking in the Cath Lab.

HIGHER-ORDER MOVEMENT

BYTE uses a constant-velocity model for state prediction. In human movement we can suspect higher-order positional derivatives to be involved [26]. Therefore, we add acceleration $\boldsymbol{a}^{(f)}$ and jerk $\boldsymbol{j}^{(f)}$ to the model. The state vector becomes

$$\boldsymbol{x}^{(f)} = \begin{bmatrix} \boldsymbol{p}^{(f)} \\ \boldsymbol{v}^{(f)} \\ \boldsymbol{a}^{(f)} \\ \boldsymbol{j}^{(f)} \end{bmatrix}, \boldsymbol{a}^{(f)} = \begin{bmatrix} ax_1^{(f)} \\ ay_1^{(f)} \\ \vdots \\ ax_{|\mathcal{K}|}^{(f)} \\ ay_{|\mathcal{K}|}^{(f)} \end{bmatrix}, \boldsymbol{j}^{(f)} = \begin{bmatrix} jx_1^{(f)} \\ jy_1^{(f)} \\ \vdots \\ jx_{|\mathcal{K}|}^{(f)} \\ jy_{|\mathcal{K}|}^{(f)} \end{bmatrix}, \tag{4.3}$$

where $\boldsymbol{p}^{(f)}$ and $\boldsymbol{v}^{(f)}$ are given by eq. (4.2) and $ax_k^{(f)}$, $ay_k^{(f)}$, $jx_k^{(f)}$, $jy_k^{(f)}$ are acceleration and jerk. In the prediction step we assume that these decrease linearly over time and update them as

$$\begin{bmatrix} \hat{\boldsymbol{a}}^{(f+1)} \\ \hat{\boldsymbol{j}}^{(f+1)} \end{bmatrix} = \left(\begin{bmatrix} \beta_a & \beta_j \\ 0 & \beta_j \end{bmatrix} \otimes I_{2|\mathcal{K}|} \right) \begin{bmatrix} \boldsymbol{a}^{(f)} \\ \boldsymbol{j}^{(f)} \end{bmatrix}, \tag{4.4}$$

where β_a and β_j are memory factors, $\otimes$ denotes the Kronecker product, and $I_x \in \mathbb{R}^{x \times x}$ is an identity matrix. Next we predict velocity and position with the 3rd-order derivative motion equations

$$\begin{bmatrix} \hat{\boldsymbol{p}}^{(f+1)} \\ \hat{\boldsymbol{v}}^{(f+1)} \end{bmatrix} = \left(\begin{bmatrix} 1 & 1 & 1/2 & 1/6 \\ 0 & 1 & 1 & 1/2 \end{bmatrix} \otimes I_{2|\mathcal{K}|} \right) \begin{bmatrix} \hat{\boldsymbol{p}}^{(f)} \\ \hat{\boldsymbol{v}}^{(f)} \\ \hat{\boldsymbol{a}}^{(f+1)} \\ \hat{\boldsymbol{j}}^{(f+1)} \end{bmatrix}. \tag{4.5}$$

We predict in these two steps to ensure that $\boldsymbol{a}^{(f)}$ and $\boldsymbol{j}^{(f)}$ have no effect on $\boldsymbol{x}^{(f+1)}$ if the respective memory factor is 0. Note that, if $\beta_a = 0$ but $\beta_j \neq 0$, jerk still causes some acceleration in eq. (4.4).

4.1.4. EXPERIMENTAL SETUP

METRICS

Detection- and tracking performance are evaluated with Detection Accuracy (DA) and Association Accuracy (AA) [27]. Higher-Order Tracking Accuracy (HOTA) $= \sqrt{\text{DA} \cdot \text{AA}}$ aggregates these metrics into a single score. We match detections to annotations as described in [27] with OKS as localisation similarity. DA, AA and HOTA are evaluated over a range of OKS thresholds from 0.5 to 0.95 with step size 0.05, and report the average as per convention [25, 27]. Additionally, we measure average algorithm speed in [fps].

WORKFLOW PHASE

Metrics are evaluated separately on each annotated workflow phase from section 4.1.1. This way we observe situational effects on pose tracking.

PARAMETERS

We use the PoseBYTE parameters from table 4.1. Optimal values for β_α and β_j are found by ranging each from 0 to 0.9 and evaluating HOTA for all workflow phases jointly. We exclude memory factors of 1 to prevent instability in the Kalman prediction step.

Table 4.1.: PoseBYTE parameters used for all experiments in section 4.1.4.

γ_{high}	γ_{kp}	σ_{high}	σ_{low}	σ_{new}	β_α	β_j	f_{mem}
0.5	0.3	0.8	0.5	0.65	0.4	0.8	50

ABLATION STUDY

We test the contribution of each PoseBYTE component on HOTA and speed. As a baseline we test bounding box tracking using IoU as object similarity. Here, we use tightest-fit bounding boxes around each pose for tracking but still evaluate metrics on keypoints for consistency. Undetected keypoints are estimated by translating and scaling the last detected pose to tightly fit the new bounding box after each prediction step. Secondly, we add pose data and OKS in the Kalman filter as described in section 4.1.3. Finally, we include the acceleration and jerk from section 4.1.3.

POSE DETECTOR

All tested pose detectors and their abbreviations in this chapter are introduced in table 4.2. All tests are carried out with AlphaP152

Table 4.2.: Pose detection models tested in section 4.1.4. Here, OpenPP30T is the only model with built-in tracking.

Model	Details
AlphaP50	ResNet50[28]+YOLOv3-SPP[29, 30]+FastPose[14]
AlphaP152	ResNet152[28]+YOLOv3-SPP[29, 30]+FastPose[14]
OpenPP16C	ShuffleNetV2K16[15, 31]+CifCaf[15]
OpenPP30C	ShuffleNetV2K30[15, 31]+CifCaf[15]
OpenPP30T	tShuffleNetV2K30[15, 31]+TrackingPose[15]

unless explicitly stated otherwise. No detectors are re-trained, i.e., the pre-trained models and code linked in the respective citations from table 4.2 are used. As the optimal values for β_a and β_j rely heavily on the pose detector, we select these to maximise HOTA separately for each detector. Other parameters are kept the same in accordance to section 4.1.4. We provide baseline tracking results from AlphaP152+Human-ReID [14] and OpenPP30T.

QUALITATIVE RESULTS

For demonstrative purposes we show example pose tracklets from AlphaP152 with Human-ReID or PoseBYTE in the 'Patient entry' phase. Frames are selected to highlight problems solved or introduced by PoseBYTE. We only show keypoints with a detection confidence of at least γ_{kp}.

4.2. RESULTS

DA is shown in table 4.3 for a range of memory factors. Scores range from 0.63 to 0.65, where a low acceleration factor seems to be preferred. Table 4.4 shows AA in a similar fasion. Here, scores range from 0.72 to 0.78 and are mostly uniform around 0.78 when β_a and β_j are lower than 0.8. Consequently, HOTA ranges from 0.67 to 0.71 with a preference for low β_a and β_j.

OpenP achieves up to 0.69 HOTA with β_a and β_j below 0.6, and achieves the best AA when $\beta_a^2 + \beta_j^2 \approx 0.55^2$. OpenPP16C prefers low β_a and β_j, and scores up to 0.64 HOTA. OpenPP30C yields up to 0.73 HOTA, when $\beta_a^2 + \beta_j^2 \leq 0.55^2$ holds. Finally, AlphaP50 scores up to 0.70 HOTA and prefers both factors between 0.4 and 0.9.

The optimal memory factors differ per workflow phase. During 'Patient entry', keeping both memory factors below 0.8 approaches a HOTA of 0.74. During 'Wrist access' making both factors 0 yields the best HOTA

Table 4.3.: DA for different acceleration- and jerk memory factors, evaluated jointly over all workflow phases.

β_a \ β_j	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.0	.64	.64	.64	.64	.64	.64	.64	.64	.65	.64
0.1	.64	.64	.64	.64	.64	.64	.64	.64	.65	.64
0.2	.64	.64	.64	.64	.64	.64	.64	.64	.64	.63
0.3	.64	.64	.64	.64	.64	.64	.64	.64	.64	.63
0.4	.64	.64	.64	.64	.64	.64	.64	.64	.64	.63
0.5	.64	.64	.64	.64	.64	.64	.64	.64	.64	.63
0.6	.64	.64	.64	.64	.64	.64	.64	.64	.64	.63
0.7	.64	.64	.64	.64	.64	.64	.64	.64	.63	.63
0.8	.64	.64	.64	.64	.64	.64	.64	.64	.63	.63
0.9	.64	.64	.64	.64	.64	.64	.64	.63	.63	.63

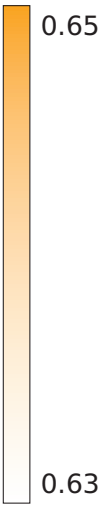
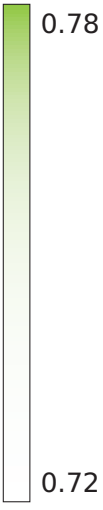


Table 4.4.: AA for different acceleration- and jerk memory factors, evaluated jointly over all workflow phases.

β_a \ β_j	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.0	.78	.78	.78	.77	.77	.78	.78	.78	.77	.77
0.1	.78	.77	.77	.77	.77	.78	.78	.77	.76	.76
0.2	.77	.78	.77	.77	.78	.78	.78	.77	.76	.75
0.3	.77	.77	.77	.77	.78	.78	.77	.77	.76	.76
0.4	.78	.78	.78	.77	.78	.78	.77	.77	.77	.76
0.5	.77	.77	.78	.78	.78	.78	.77	.76	.77	.76
0.6	.78	.78	.78	.78	.78	.78	.77	.76	.77	.75
0.7	.78	.78	.78	.77	.78	.77	.76	.76	.75	.75
0.8	.77	.77	.78	.77	.76	.76	.77	.77	.75	.75
0.9	.77	.77	.76	.76	.77	.77	.76	.76	.75	.72



of 0.70. The 'Ultrasound' phase prefers factors of $\beta_a^2 + \beta_j^2 \approx 0.6^2$ for a HOTA of 0.75. 'X-Ray' yields 0.71 HOTA for $\beta_j \approx 0.9 - 0.4\beta_a$. During 'Wound closure' the best DA of 0.56 is achieved for $\beta_a^2 + \beta_j^2 \approx 0.8$, and the best AA of 0.76 for $\beta_a^2 + \beta_j^2 \approx 0.55$.

Table 4.5.: PoseBYTE HOTA and speed in fps per workflow phase with the parameters from table 4.1 after each addition from section 4.1.3

OKS	β_a	β_j	**Patient entry**		**Wrist access**		**Ultrasound**		**X-Ray**		**Wound closure**		**Total**	
			HOTA	fps	HOTA	fps	HOTA	fps	HOTA	fps	HOTA	fps	HOTA	fps
			.74	21.7	.69	27.5	.75	18.6	.66	20.8	.61	24.4	.70	22.2
✓			.73	21.8	.71	27.8	.75	18.1	.68	22.8	.64	25.3	.71	22.7
✓	.7		.73	21.5	.69	27.5	.75	18.2	.68	22.9	.64	24.4	.70	22.5
✓	.2	0.5	.73	21.3	.69	26.7	.75	18.0	.68	23.2	.65	25.2	.71	22.5

Table 4.5 shows HOTA and inference speed per added PoseBYTE component. Tracking poses instead of bounding boxes increases HOTA by 0 pp to 3 pp depending on the phase. An exception is the 'Patient entry' phase, on which HOTA decreases by 1 pp. Adding β_a keeps results mostly the same, where HOTA decreases by 2 pp during 'Wrist access'. The addition of β_j increases HOTA by 1 pp on the 'Wound closure' phase. Speed changes per added component seem negligible, where the largest observed change on all phases jointly is 0.5 fps. We observe speed differences per workflow phase, where 'Ultrasound' yields the lowest speeds of 18.0 fps to 18.6 fps and 'Wrist access' the highest of 26.7 fps to 27.8 fps.

For OpenPP16C, which achieves a HOTA score of 0.44 with BYTE, the addition of OKS yields a HOTA gain of 20 pp. OpenPP30C sees a similar increase from 0.58 to 0.73. OpenP gains 5 pp with OKS over 0.64 HOTA with BYTE.

We show results for all considered pose detectors and trackers in table 4.6. The best HOTA of 0.73 is achieved by OpenPP30C+PoseBYTE, with a speed of 5.0 fps. It is closely followed with 0.71 HOTA by AlphaP152+PoseBYTE—the fastest model at 22.5 fps. AlphaP50+PoseBYTE comes close with 0.70 HOTA at 19.1 fps. The lowest HOTA and speed come from OpenPP30T: 0.53 at 4.3 fps

AlphaP and OpenPP achieve higher HOTA and speed with PoseBYTE than with their own trackers—Human-ReID and TrackingPose. For AlphaP the HOTA improvement is up to 6 pp whereas for OpenPP it is 20 pp. Looking per phase, Human-ReID outperforms PoseBYTE by up to 5 pp during 'Wrist access', 'Ultrasound' and 'X-Ray'. During 'Patient entry', PoseBYTE outperforms Human-ReID with up to 19 pp. OpenP yields worse HOTA and speed than AlphaP152 with PoseBYTE on all workflow

Table 4.6.: HOTA and speed in fps of PoseBYTE and other trackers

Model	Patient entry		Wrist access		Ultrasound		X-Ray		Wound closure		Total	
	HOTA	fps	HOTA	fps	HOTA	fps	HOTA	fps	HOTA	fps	HOTA	fps
AlphaP50+ Human-ReID	0.55	12.4	0.69	15.8	0.74	11.0	0.73	14.4	0.52	16.1	.65	13.7
AlphaP50+ PoseBYTE(ours)	0.74	17.1	0.68	25.8	0.72	14.8	0.70	20.3	0.65	20.6	.70	19.1
AlphaP152+ Human-ReID	0.55	12.2	0.70	15.4	0.76	11.0	0.73	14.2	0.53	16.0	.65	13.5
AlphaP152+ PoseBYTE(ours)	0.73	21.3	0.69	26.7	0.75	18.0	0.68	23.2	0.65	25.2	.71	22.5
OpenPP16C+ PoseBYTE(ours)	0.71	9.4	0.54	9.8	0.62	8.8	0.65	9.7	0.64	9.8	.64	9.5
OpenPP30T	0.60	4.2	0.47	4.3	0.53	4.2	0.54	4.3	0.51	4.3	.53	4.3
OpenPP30C+ PoseBYTE(ours)	0.77	5.0	0.68	5.1	0.74	4.8	0.75	5.1	0.69	5.1	.73	5.0
OpenP+ PoseBYTE(ours)	0.71	10.1	0.60	10.1	0.75	9.9	0.74	9.8	0.62	10.2	.69	10.0

phases except ‘Ultrasound’ and ‘X-Ray’. For OpenPP16C there are no such exceptions. AlphaP152+PoseBYTE yields HOTA differences per phase of up to 10 pp, which is 23 pp with Human-ReID. This difference is present but less pronounced for OpenPP30 with 9 pp and 13 pp.

Figure 4.2 shows qualitative results of Human-ReID and PoseBYTE during ‘Patient entry’ with the AlphaP152 detector. Each row shows a different frame in chronological order, where the top row comes first and bottom row last. Time intervals between rows are not constant. Each pose shows an integer tracking ID and the detection confidence score between 0 and 1.

At the start of the procedure, Human-ReID sees all persons earlier than PoseBYTE. Shortly after, PoseBYTE catches up and sees the same persons. Between the second and third timesteps, one assistant passes in front of the patient and another walks behind the infusion bags. Here, an identity swap occurs with Human-ReID between the patient and the first assistant, but not with PoseBYTE. Both trackers lose the second assistant, after which Human-ReID wrongly assigns a previously seen ID and PoseBYTE assigns a new one. Only Human-ReID sees the third assistant in the lower-left corner. In the fourth row, Human-ReID re-assigned the first two assistants their initial IDs. A duplicate pose can be seen in the patient, which is now assigned both their original ID and that of the third assistant in the corner. PoseBYTE is still tracking the two assistants, but has assigned a new ID to the patient after an assistant passed them in front. In the last row, neither Human-ReID nor

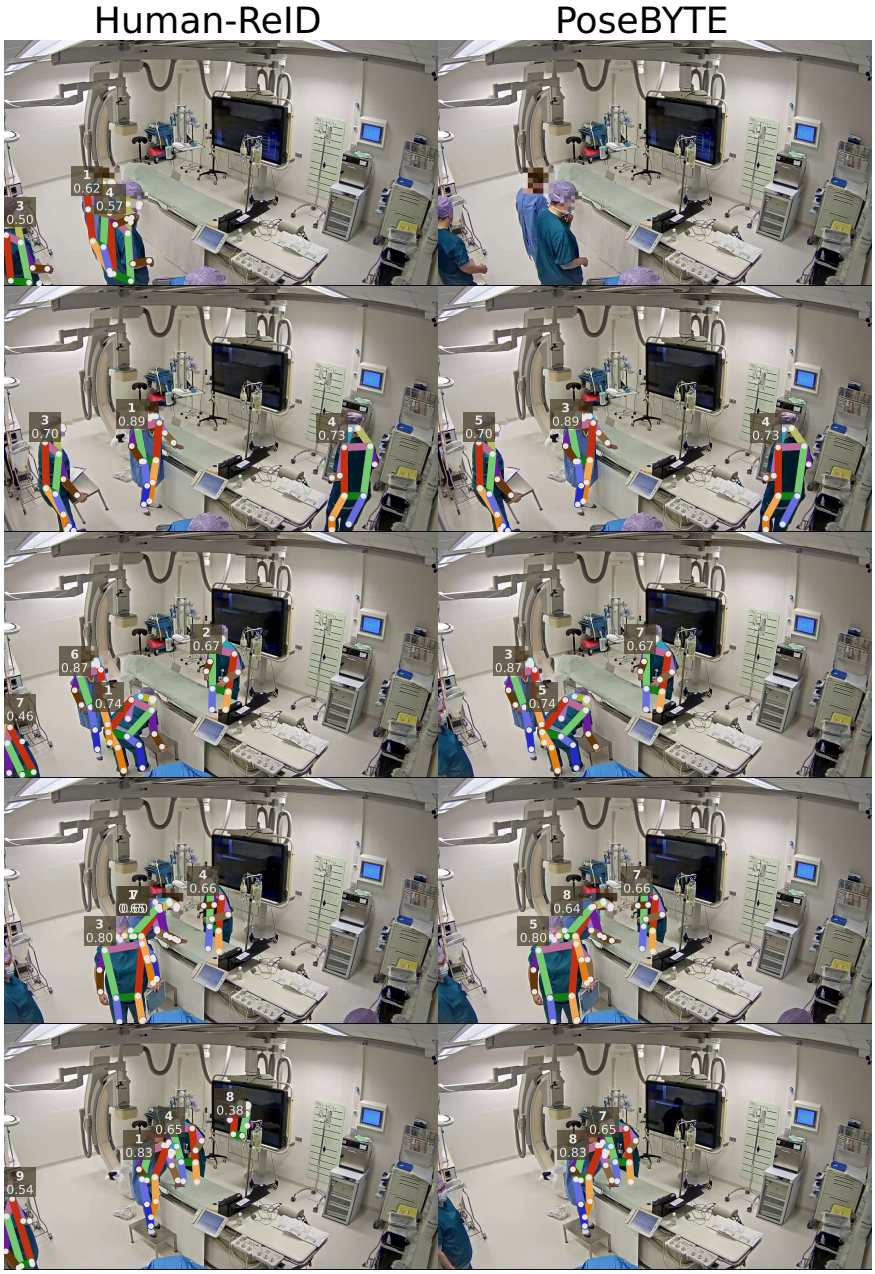


Figure 4.2.: Qualitative results of Human-ReID and PoseBYTE during the ‘Patient entry’ phase, where poses are detected with AlphaP152. Rows show different timeframes in chronological order from top to bottom, with varying intervals.

PoseBYTE has lost or swapped any IDs. Here, Human-ReID sees a pose in the reflection of the monitor, which PoseBYTE ignores because of the tracklet confirmation step inherent to BYTE.

4.3. DISCUSSION

In this work we adapted BYTE for pose tracking in the Cath Lab and compared the resulting method to pose trackers from literature.

During the 'Patient entry' phase, Human-ReID and TrackingPose yield HOTA scores of up to 0.60. We observe in videos that Human-ReID is prone to identity swaps, which could be due to it relying on visual clues of similarly-dressed personnel. It occasionally detects duplicate poses, which slip past the non-maximum suppression designed to solve this very problem [14]. TrackingPose sometimes merges poses that are close to each other, possibly because of its multi-frame pose construction creating more opportunity to do so. It also tends to miss visible keypoints in partially occluded poses, which Human-ReID solves possibly by imposing a prior through bounding box detection. The many (4 to 6) visible persons during 'Patient entry' could amplify these issues. PoseBYTE uses no visual features and does tracking and detection separately, which could contribute to it performing up to 18 pp HOTA better on this phase. However, this tracker does tend to lose persons quickly during occlusions of more than a few frames.

With the 'Ultrasound' phase containing 4 to 5 people, one could expect the same problems to occur. Although TrackingPose performs similarly here, Human-ReID does better with up to 0.76 HOTA—1 pp higher than PoseBYTE. A difference between this phase and 'Patient entry' is that, although people occlude each other in both, they walk a lot during 'Patient entry' and stay in place during 'Ultrasound'. Their close vicinity causes TrackingPose the same problems as before, whilst their stillness could be allowing Human-ReID to track more accurately based on position. The same is visible in the other low-movement phases 'Wrist access' and 'X-Ray', which both have only 2 to 3 persons in the room besides the—rarely visible—patient, simplifying the tracking problem.

Modelling acceleration and jerk yielded little HOTA improvement for any tested detector. It yielded its largest HOTA improvement of 5 pp when using the AlphaP50 detector in the 'Wound closure' phase. Different combinations of detector and phase yield different optimal memory factors. All in all, the benefit of including higher-order movement seems negligible.

We aim to provide a tracker that works reliably throughout a procedure. Even though PoseBYTE does not always perform better than the benchmark set by the state-of-the-art, its HOTA varies much less between workflow phases. For workflow analysis the most important

phases to analyse through poses are ‘Patient entry’ and ‘Wound closure’, as during other phases the system logs provide an alternative source of workflow information. During these phases, PoseBYTE delivers a HOTA improvement of 12 pp to 19 pp with respect to the benchmark.

PoseBYTE speed roughly decreases with the number of people in the room when the AlphaP detector is used. This effect is much less pronounced, if at all, with the OpenPP and OpenP detectors. The same can be observed with Human-ReID and TrackingPose, suggesting that the slowdown occurs in the detection- and not the tracking stage. In either case, PoseBYTE achieves higher speeds than the benchmark trackers in all situations, making real-time applications more viable.

For the purpose of workflow analysis, it is important to have a reliable information source throughout a procedure. PoseBYTE fits this description well, as it achieved the lowest HOTA spread of all tested trackers over the tested CAG workflow phases. Especially during ‘Patient entry’, where few other sources of workflow information are available, PoseBYTE improves on the state-of-the-art. Whether its overall HOTA score of 0.71 is high enough will depend on what information one wants to obtain. It will suffice for measuring estimate positions and short-term motion of people in the Cath Lab, which can already be indicative of workflow. However, the results might not be good enough for analysis of fine-grained long-term movement and gestures. Here, the tendency of PoseBYTE to lose tracklets after occlusion could interfere. One can mitigate the effect of inaccurate motion model predictions by excluding keypoints with a confidence below γ_{kp}.

We did not test for optimal values of PoseBYTE parameters other than β_a and β_j, and even those latter two were tested only over a limited set of values. For reference, a memory factor of 0.9 per frame amounts to a memory of only $0.9^{25} = 0.072$ per second. The used movement model assumes keypoints to move independently of each other, causing anatomically unrealistic predictions over time. This could explain why PoseBYTE still has trouble reidentifying poses after occlusions of some frames.

In future work more memory factors in the range $[0.9, 1)$ could be tested, in addition to finding optimal values for other parameters. A memory factor for velocity could be included, as we observe movements in the Cath Lab to often span short distances. Alternatively a different model could be used, built specifically for human motion prediction [32]. Finally, the integration of multiple camera views could be investigated as in [33–35].

PoseBYTE yields higher HOTA and speed in the Cath Lab with greater stability between different situations than the tested pose trackers from literature. With a HOTA score of 0.71 at 22.5 fps, it is a suitable method for short-term real-time pose tracking for workflow analysis in the Cath Lab.

4.4. CONCLUSION

We adapted BYTE for pose tracking in the Cath Lab without relying on visual clues. The algorithm was evaluated in terms of HOTA and speed on five annotated CAG workflow phases before, during, and after procedures. PoseBYTE shows stable performance across workflow phases and outperforms the current state of the art in terms of HOTA and speed. The improvement is most apparent when the patient enters the room, which is also the least trivial situation for tracking.

REFERENCES

[1] R. M. Butler, T. S. Vijfvinkel, E. Frassini, S. van Riel, C. Bachvarov, J. Constandse, M. van der Elst, J. J. van den Dobbelsteen, and B. H. W. Hendriks. “2D Human Pose Tracking in the Cardiac Catheterisation Laboratory with BYTE”. In: *Med. Eng. Phys.* 135 (Jan. 2025), p. 104270. doi: 10.1016/j.medengphy.2024.104270.

[2] K. N. Timoh, A. Huaulme, K. Cleary, M. A. Zaheer, V. Lavoué, D. Donoho, and P. Jannin. “A Systematic Review of Annotation for Surgical Process Model Analysis in Minimally Invasive Surgery based on Video”. In: *Surg. Endosc.* 37 (May 2023), pp. 4298–4314. doi: 10.1007/s00464-023-10041-w.

[3] A. M. Schouten, S. M. Flipse, K. E. van Nieuwenhuizen, F. W. Jansen, A. C. van der Eijk, and J. J. van den Dobbelsteen. “Operating Room Performance Optimization Metrics: a Systematic Review”. In: *J. Med. Syst.* 47.1 (Dec. 2023), p. 19. doi: 10.1007/s10916-023-01912-9.

[4] F. Lalys and P. Jannin. “Surgical Process Modelling: a Review”. In: *Int. J. Comput. Assist. Radiol. Surg.* 9 (May 2014), pp. 495–511. doi: 10.1007/s11548-013-0940-5.

[5] Mayo Clinic Staff. *Coronary Angiogram*. Dec. 2023. url: https://www.mayoclinic.org/tests-procedures/coronary-angiogram/about/pac-20384904.

[6] G. W. Reed, S. Hantz, R. Cunningham, A. Krishnaswamy, S. G. Ellis, U. Khot, J. Rak, and S. R. Kapadia. “Operational Efficiency and Productivity Improvement Initiatives in a Large Cardiac Catheterization Laboratory”. In: *JACC: Cardiovasc. Interv.* 11.4 (Feb. 2018), pp. 329–338. doi: 10.1016/j.jcin.2017.09.025.

[7] C. R. Garrow, K.-F. Kowalewski, L. Li, M. Wagner, M. W. Schmidt, S. Engelhardt, D. A. Hashimoto, H. G. Kenngott, S. Bodenstedt, S. Speidel, B. P. Müller-Stich, and F. Nickel. “Machine Learning for Surgical Phase Recognition: A Systematic Review”. In: *Annal. Surg.* 273.4 (Apr. 2021), pp. 684–693. doi: 10.1097/SLA.0000000000004425.

[8] M. Berlet, T. Vogel, D. Ostler, T. Czempiel, M. Kähler, S. Brunner, H. Feussner, D. Wilhelm, and M. Kranzfelder. “Surgical Reporting for Laparoscopic Cholecystectomy based on Phase Annotation by a Convolutional Neural Network (CNN) and the Phenomenon of Phase Flickering: a Proof of Concept”. In: *Int. J. Comput. Assist. Radiol. Surg.* 17 (Nov. 2022), pp. 1991–1999. doi: 10.1007/s11548-022-02680-6.

[9] I. Aksamentov, A. P. Twinanda, D. Mutter, J. Marescaux, and N. Padoy. “Deep Neural Networks Predict Remaining Surgery Duration from Cholecystectomy Videos”. In: *Med. Image Comput. Comput.-Assist. Interv.* Springer, Sept. 2017, pp. 586–593. doi: 10.1007/978-3-319-66185-8_66.

[10] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, M. Hashizume, D. Katic, H. Kenngott, M. Kranzfelder, A. Malpani, K. März, T. Neumuth, N. Padoy, C. Pugh, N. Schoch, D. Stoyanov, R. Taylor, M. Wagner, G. D. Hager, and P. Jannin. “Surgical Data Science for Next-generation Interventions”. In: *Nat. Biomed. Eng.* 1 (Sept. 2017), pp. 691–696. doi: 10.1038/s41551-017-0132-7.

[11] G. Saleem, U. I. Bajwa, and R. H. Raza. “Toward Human Activity Recognition: a Survey”. In: *Neural Comput. Appl.* 35 (Feb. 2023), pp. 4145–4182. doi: 10.1007/s00521-022-07937-4.

[12] H.-C. Nguyen, T.-H. Nguyen, and V.-H. Scherer Rafałand Le. “Deep Learning for Human Activity Recognition on 3D Human Skeleton: Survey and Comparative Study”. In: *Sens.* 23.11 (May 2023), p. 5121. doi: 10.3390/s23115121.

[13] C. Wang and J. Yan. “A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition”. In: *IEEE Access* 11 (June 2023), pp. 53880–53898. doi: 10.1109/ACCESS.2023.3282311.

[14] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu. “AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 45.6 (June 2023), pp. 7157–7173. doi: 10.1109/TPAMI.2022.3222784.

[15] S. Kreiss, L. Bertoni, and A. Alahi. “OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association”. In: *IEEE Trans. Intell. Transp. Syst.* 23.8 (Aug. 2022), pp. 13498–13511. doi: 10.1109/TITS.2021.3124981.

[16] M. Wang, J. Tighe, and D. Modolo. “Combining Detection and Tracking for Human Pose Estimation in Videos”. In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2020, pp. 11085–11093. doi: 10.1109/CVPR42600.2020.01110.

[17] Z. Wang, H. Zhao, Y.-L. Li, S. Wang, P. H. Torr, and L. Bertinetto. “Do Different Tracking Tasks Require Different Appearance Models?” In: *Adv. Neural Inf. Process. Syst.* Curran Associates Inc., Dec. 2021, pp. 726–738.

[18] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele. “PoseTrack: A Benchmark for Human Pose Estimation and Tracking”. In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2018, pp. 5167–5176. doi: 10.1109/CVPR.2018.00542.

[19] Y. Zhang, P. Sun, y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang. “ByteTrack: Multi-object Tracking by Associating Every Detection Box”. In: *Eur. Conf. Comput. Vis.* Springer, Oct. 2022, pp. 1–21. doi: 10.1007/978-3-031-20047-2_1.

[20] CVAT.ai Corporation. *Computer Vision Annotation Tool (CVAT)*. url: https://www.cvat.ai.

[21] R. E. Kalman. “A New Approach to Linear Filtering and Prediction Problems”. In: *J. Basic Eng.* 82.1 (Mar. 1960), pp. 35–45. doi: 10.1115/1.3662552.

[22] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye. “Object Detection in 20 Years: A Survey”. In: *Proc. IEEE* 111.3 (Jan. 2023), pp. 257–276. doi: 10.1109/JPROC.2023.3238524.

[23] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. da Silva. “A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit”. In: *Electron.* 10.3 (Jan. 2021), p. 279. doi: 10.3390/electronics10030279.

[24] H. W. Kuhn. “Variants of the Hungarian Method for Assignment Problems”. In: *Nav. Res. Logist. Q.* 03.04 (Dec. 1956), pp. 253–258. doi: 10.1002/nav.3800030404.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. “Microsoft COCO: Common Objects in Context”. In: *Eur. Conf. Comput. Vis.* Springer, Sept. 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[26] R. Sers, S. Forrester, M. Zecca, S. Ward, and E. Moss. “Objective Assessment of Surgeon Kinematics during Simulated Laparoscopic Surgery: a Preliminary Evaluation of the Effect of High Body Mass Index Models”. In: *Int. J. Comput. Assist. Radiol. Surg.* 17.1 (Jan. 2022), pp. 75–83. doi: 10.1007/s11548-021-02455-5.

[27] J. Luiten, A. Ošep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe. “HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking”. In: *Int. J. Comput. Vis.* 129.2 (Feb. 2021), pp. 548–578. doi: 10.1007/s11263-020-01375-2.

[28] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *Proc. 29th IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[29] J. Redmon and A. Farhadi. *YOLOv3: An Incremental Improvement*. Apr. 2018. doi: 10.48550/arXiv.1804.02767. arXiv: 1804.02767v1 [cs.CV].

[30] K. He, X. Zhang, S. Ren, and J. Sun. “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 37.9 (Sept. 2015), pp. 1904–1916. doi: 10.1109/TPAMI.2015.2389824.

[31] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. “ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design”. In: *Proc. 15th Eur. Conf. Comput. Vis.* Springer, Sept. 2018, pp. 122–138. doi: 10.1007/978-3-030-01264-9_8.

[32] J. Martinez, M. J. Black, and J. Romero. “On Human Motion Prediction Using Recurrent Neural Networks”. In: *IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, July 2017, pp. 4674–4683. doi: 10.1109/CVPR.2017.497.

[33] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin, and N. Padoy. “Articulated Clinician Detection using 3D Pictorial Structures on RGB-D Data”. In: *Med. Image Anal.* 35 (Jan. 2017), pp. 215–224. doi: 10.1016/j.media.2016.07.001.

[34] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin, and N. Padoy. “A Multi-View RGB-D Approach for Human Pose Estimation in Operating Rooms”. In: *IEEE Winter Conf. Appl. Comput. Vis.* IEEE, Mar. 2017, pp. 363–372. doi: 10.1109/WACV.2017.47.

[35] A. Kadkhodamohammadi and N. Padoy. “A Generalizable Approach for Multi-View 3D Human Pose Regression”. In: *Mach. Vis. Appl.* 32 (Oct. 2021), p. 6. doi: 10.1007/s00138-020-01120-2.

II

WORKFLOW ANALYSIS


5

Workflow phase estimation from 2D human motion

Previous chapters refined 2D human pose tracking in the cardiac catheterisation laboratory. Human pose tracklets are a popular input feature to human- and group action recognition algorithms. Existing methods vary in terms of accuracy and interpretability. Workflow phases during cardiac angiograms are a form of group activity. As an explorative workflow study, we investigate the relation between human motion and workflow phases in a transparent way. Since neural network reasoning is difficult to interpret, we build a classifier on handcrafted features instead. History vectors measure keypoint motion and distances over time. Using expectation maximisation, we build mixture models to represent each workflow phase. Given a new history vector measurement, each mixture model produces a score for it indicating the respective workflow phase. In addition, a workflow phase prior was imposed based on procedure duration. The trained model yielded near-uniform workflow phase probabilities, regardless of the ground truth. It was biased towards phases that appeared more often, yielden better-than-uniform accuracies up to 0.39. The flawed results suggested that different keypoints contribute differently towards the prediction. For example, the distance between persons seemed to indicate procedure duration. Although the proposed method did not yield accurate workflow phase classifications, provided insights can be used in future algorithm design. Associating information between different keypoints, which the proposed model lacked, is seemingly important.

Operating rooms (ORs) account for a large fraction of the hospital budget. Personnel shortages and rising procedure costs further complicate the situation in modern healthcare [1–3]. It is more important than ever to make efficient use of operating rooms to alleviate these circumstances. A first step towards more efficient use is to obtain insights into current peri-operative workflows, and determine possible best practices or inefficiencies [4, 5]. This could be done through manual observation, but this is labour-intensive and does not scale to other hospitals. In this chapter, we explore computer-assisted methods for automated observation [6, 7]. This approach requires fewer resources than manual annotation, but is technically challenging [8].

The Cardiac Catheterisation Laboratory (Cath Lab) is a specialised OR for minimally invasive cardiac procedures. The Cath Lab is equipped for its purpose with a 'C-Arm'-mounted X-Ray device, a monitor, a mobile operating table and a radiation shield. During procedures, a cardiologist, scrub nurse, several assistants, and occasional spectators are present. The cardiologist and scrub nurse control the devices using a control panel, offering physical buttons, a touch screen, and/or foot pedals. Patient vitals and X-ray images are viewed on the monitor in real-time. For radiation-safety, everyone but the cardiologist and scrub nurse—and possibly a spectator—retreat to the adjacent control room during the procedure. Here, they receive the same information as the Cath Lab monitor on a separate display, and have a clear view into the room through a radiation-proof window. Those who remain in the room wear lead aprons, glasses and collars, and can position the lead shield to receive as little exposure as possible.

One diagnostic procedure performed in the Cath Lab is the Cardiac Angiogram (CAG) [9]. A catheter is guided through the wrist or groin to administer a contrast fluid into the heart. The contrast fluid reveals blood flow through the cardiac arteries in X-Ray footage, showing deficiencies. From a workflow perspective, the CAG is relatively straightforward and standardised. This makes it a suitable target for explorative workflow study.

Workflow phases can be extracted from procedure video footage. This approach has clear parallels to Human Action Recognition (HAR) [10]. Human actions can be divided into different levels of complexity and collaboration. from this perspective, a workflow phase is a 'group activity' [11].

Published datasets with annotated group activities were recorded in public areas [11]. These usually consist of short videos lasting up to 30 s, with a single activity label each. To our knowledge, no public dataset contains footage from medical procedures. In such procedures, which can last hours, The presence of phase transitions demands multiple classifications over time.

Human pose tracklets are a popular feature for HAR [12, 13]. A pose

is defined as a collection of keypoints like in [14]. Pose tracklets contain a unique identifier (ID) to reidentify individuals between video frames. The set of all keypoints in a range of video frames $\mathcal{F}$ can be defined as

$$\left\{k_{p,K}^{(f)} : p \in \mathcal{P}, K \in \mathcal{K}, f \in \mathcal{F}\right\}, \tag{5.1}$$

$$k_{p,K}^{(f)} \in \mathbb{R}^2, \tag{5.2}$$

$$\mathcal{F} = \{f : f = 1, \cdots, F\}, \tag{5.3}$$

where $\mathcal{P}$ is the set of all individuals, $\mathcal{K}$ is a predefined set of keypoint classes that form a pose like in [14], and F is the number of frames. $k_{p,K}^{(f)}$ contains image-space cartesian coordinates in pixels (px). One can obtain e.g. a single pose detection by fixing p and f and varying K, a keypoint tracklet by fixing p and K and varying f, or a pose tracklet by fixing only p. The process of pose tracking aims to find a set of detections

$$\left\{d_{p,K}^{(f)} : p \in \mathcal{P}, K \in \mathcal{K}, f \in \mathcal{F}\right\}, \tag{5.4}$$

$$d_{p,K}^{(f)} \approx k_{p,K}^{(f)}. \tag{5.5}$$

Since in practice usually $d_{p,K}^{(f)} \neq k_{p,K}^{(f)}$, a pose tracker additionally produces a detection confidence

$$c_{p,K}^{(f)} \in [0, 1] \tag{5.6}$$

to accompany each $d_{p,K}^{(f)}$.

As poses naturally translate to partially connected graphs, modern works often use Graph Convolutional Networks (GCNs) for HAR from human poses [10–13]. Other popular approaches include long short-term memory (LSTM) and Transformer models, which excel in finding temporal relations. Yokoyama et al. [15] explores workflow phase recognition from real OR footage, using a LSTM network that provides a classification per frame. Neural networks—which all methods mentioned are examples of—provide state-of-the-art inference speed and accuracy. A drawback is their 'black box' nature, which makes the reasoning behind a classification hard to interpret without the proper tools [16].

In this work, we aim to classify CAG workflow phases from 2D human tracklets in an interpretable manner. The tracklets are obtained using a neural network. The velocity history vector from [17] is an interpretable, handcrafted feature that was succesfully used for HAR. We introduce the history vector: a generalisation of the velocity history vector that can look at the historical positioning of any keypoint with respect to itself or any other keypoint. By extracting history vectors from detected tracklets, we look at the pose, position and movement of individuals—with respect to themselves or each other—as

an indicator of the workflow phase. We modify the classification process from [17] to discriminate between keypoint classes, and correct for detection confidence. We demonstrate how the handcrafted features make transparent how different keypoints and relations contribute to classification. The algorithm is trained on video recordings of 7 real CAG procedures, in which workflow phases were labelled each second.

Section 5.1 describes the dataset and annotations, pose detection- and tracking algorithm, and the classification and training process. The section finishes with a description of performed experiments, the results of which are discussed in section 5.2. Section 5.3 interprets the results and their implications. Section 5.4 concludes the work.

5.1. PHASE ESTIMATION

5.1.1. VIDEOS AND ANNOTATIONS

Figure 5.1.: The Cath Lab viewpoint evaluated in this work.

Seven real Cardiac Angiogram (CAG) procedures were recorded in the Cath Lab of the Reinier de Graaf hospital (RdGG) in Delft, NL. Recordings were shot in 1920 px $\times$ 1088 px with 25 frames per second using multiple Axis M1125 cameras. In this work we use the viewpoint shown in fig. 5.1, which was experimentally verified to yield the best pose tracking results whilst providing a clear view on workflow activities. The low-level CAG workflow phases described in table 5.1 were annotated in Noldus Oberserver XT [18] by two medical doctors studying workflow in the RdGG. Figure 5.2 shows the representation of each phase in the dataset.

Table 5.1.: Annotated low-level CAG workflow phases, and grouping into high-level phases.

Low-level phases		High-level phases	
Phase	Description	Phase	Description
A	Preparation	𝔸	Preparation
B	Patient arrival		
C	Preparation with patient		
D	Endovascular access	𝔹	Recording 1
E	Guiding catheter 1		
F	Recording artery 1		
Fa	Extra catheter during F		
G	Guiding catheter 2	ℂ	Recording 2
H	Recording artery 2		
Ha	Extra catheter during H		
I	Preparing wound closure		
J	Wound closure		
K	Cleaning (patient present)	𝔻	Cleaning
L	Cleaning (patient absent)		

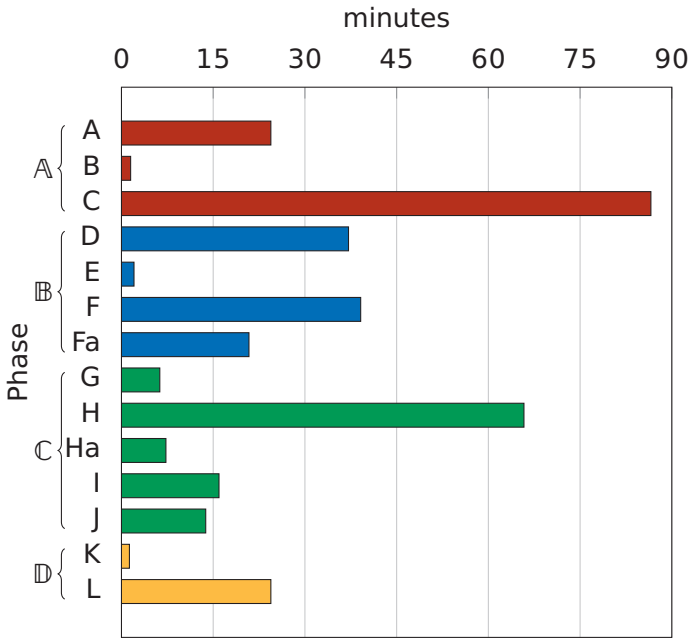


Figure 5.2.: The number of hours per workflow phase over the included dataset.

Large differences can be seen, which could make the recognition of some phases more difficult than others when trained on this dataset.

5.1.2. POSE ESTIMATION

FastPose [19]—pretrained on COCO [14]—was used to detect human poses. YOLOv3-SPP [20, 21] and ResNet152 [22] supplied the required bounding boxes and feature maps, respectively. Poses on different frames were combined into tracklets and refined with PoseBYTE [23]. Keypoint detections with a confidence below 0.3 were considered undetected. Tracklet gaps were filled with bilinear spline interpolation if the gap was neighboured by two detected keypoints.

5.1.3. WINDOWING

We divide an input video into a set of overlapping windows. A new window starts every S frames. Each window has a length of F frames. We allow the last window to be shorter if the video length is not divisible by F.

A single workflow phase is predicted per window, as explained in the following subsections. Pose tracklets are sampled and zero padded to fit the window and serve as the input sample. The workflow phase annotated on the last window frame serves as ground truth. Selecting a larger F provides more context during prediction, but increases computational complexity.

5.1.4. HISTORY VECTORS

A history vector [17] contains the binned position of a keypoint detection $d^{(f)}_{p,K}$ with respect to a reference $d^{(f_0)}_{p_0,K_0}$ over a range of frames f. The reference is produced by a function $r(p,K,f) = \{p_0, K_0, f_0\}$, which can be designed to capture different aspects of motion as demonstrated in section 5.1.4. Assuming that a single p, K, and r are selected and kept fixed, f_0, p_0 and K_0 are implied and notation simplifies to

$$d^{(f)} := d^{(f)}_{p,K}, \tag{5.7}$$

$$d^{(f)}_0 := d^{(f_0)}_{p_0,K_0} \tag{5.8}$$

on frame f. As $d^{(f)}_0$ is implied by r and $d^{(f)}$, we omit it in future equations. $c^{(f)}_{p,K}$ and $c^{(f_0)}_{p_0,K_0}$ are implied by $d^{(f)}$ and $d^{(f)}_0$, and are therefore also omitted.

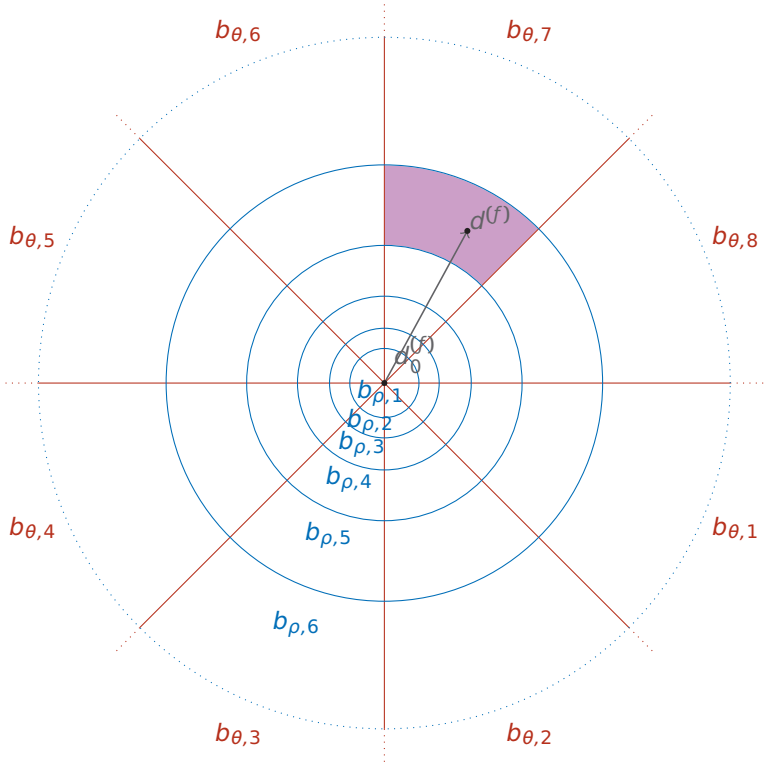


Figure 5.3.: History vector bins for $B_\rho = 6$, $B_\theta = 8$, $\rho_{\text{first}} = 0.5$ and $\rho_{\text{max}} = 5$. A displacement between two keypoints is assigned to a single bin.

VELOCITY HISTORY VECTORS

Reference [17] introduces the velocity history vector, which represents the motion of a single tracked keypoint. Here, $r(p, K, f) = \{p, K, f - \Delta f\}$ compares a keypoint with its own historic position, where a larger $\Delta f \in \mathbb{N}^+$ captures longer-term motion.

INTRA-DISTANCE HISTORY VECTORS

We introduce the intra-distance history vector to capture relative keypoint positions within a pose, with $r(p, K, f) = \{p, K_0, f\}$ for an arbitrary $K_0 \in \mathcal{K}, K_0 \neq K$. This representation captures relations within a pose.

INTER-DISTANCE HISTORY VECTORS

Where the previous subsections focus on movement within poses, positioning of individuals with respect to each other also describes workflow. For example, the cardiologist and lab assistant are standing close together during recording phases whereas many people walk around during preparation. To this end, we design the inter-distance history vector as $r(p, K, f) = \{p_0, K, f\}$ for an arbitrary reference pose $p_0 \in \mathcal{P}, p_0 \neq p$.

BINNING

Magnitude and orientation of displacement $d^{(f)} - d_0^{(f)}$ are assigned to bins from predefined sets

$$\mathcal{B}_\rho = \left\{ b_{\rho,n} : n \leq B_\rho, n \in \mathbb{N}^+ \right\}, \tag{5.9}$$

$$\mathcal{B}_\theta = \left\{ b_{\theta,m} : m \leq B_\theta, m \in \mathbb{N}^+ \right\}, \tag{5.10}$$

respectively, where B_ρ and B_θ are the number of magnitude- and orientation bins. $\mathcal{B}_\rho$ and $\mathcal{B}_\theta$ combine into a larger set

$$\begin{aligned} \mathcal{B} = \big\{ b_k : b_k = b_{\rho,n} \cap b_{\theta,m}, \\ b_{\rho,n} \in \mathcal{B}_\rho, b_{\theta,m} \in \mathcal{B}_\theta, \\ k = (n-1)B_\theta + m \big\}. \end{aligned} \tag{5.11}$$

Figure 5.3 shows an example grid of bins, with one binned value. Orientation bins are spaced uniformly. To retain detail in fine-grained movement, boundaries between magnitude bins $b_{\rho,n} \in \mathcal{B}_\rho$ are spaced logaritmically using the natural logarithm, between the desired endpoints ρ_{first} and ρ_{last}. $b_{\rho,1}$ extends from 0 to ρ_{first}, and b_{ρ,B_ρ} ends at ρ_{last}. Any magnitudes beyond ρ_{last} are still assigned to b_{ρ,B_ρ}.

Binning produces a history vector

$$\boldsymbol{h} = \begin{bmatrix} h^{(1)} & \cdots & h^{(F)} \end{bmatrix}^T, \tag{5.12}$$

where $h^{(f)}$ contains the bin assignment on window frame f.

5.1.5. MARKOV MODELS

Reference [17] makes the assumptions that

1. each history vector is generated by a Markov model, and
2. each generating Markov model is unique to one activity (or workflow phase, in our situation).

A Markov model M provides history vector sample probabilities

$$p_{h^{(f)}|M}(b|m) \tag{5.13}$$

and transition probabilities

$$p_{h^{(f)}|h^{(f-1)},M}(b|b_0, m). \tag{5.14}$$

We estimate the probability that a history vector was detected given Markov model M as

$$p_{\boldsymbol{h}|M}(\boldsymbol{b}|m) = p_{h^{(1)}|M}(b|m) \prod_{f=2}^{F} p_{h^{(f)}|h^{(f-1)},M}(b|b_0, m). \tag{5.15}$$

There may be frames on which keypoints were not detected, leaving gaps in their tracklets and thus $\boldsymbol{h}$. If missing detections separate $\boldsymbol{h}$ into a set of history subvectors $\mathcal{H}$, we write

$$p_{\boldsymbol{h}|M}(\boldsymbol{b}|m) = \prod_{\hat{\boldsymbol{h}} \in \mathcal{H}} p_{\hat{\boldsymbol{h}}|M}(\boldsymbol{b}|m), \tag{5.16}$$

calculating each $p_{\hat{\boldsymbol{h}}|M}(\boldsymbol{b}|m)$ as in eq. (5.15).

$p_{\boldsymbol{h}|M}(\boldsymbol{b}|m)$ decreases as F grows. Thus, long sequence lengths yield lower probabilities than short ones. Therefore, we normalise eq. (5.16) as

$$p^{1/\hat{F}}_{\boldsymbol{h}|M}(\boldsymbol{b}|m), \tag{5.17}$$

where $\hat{F}$ is the total number of frames in $\mathcal{H}$.

In numerical calculation, rounding errors cause $p_{\boldsymbol{h}|M}(\boldsymbol{b}|m)$ to become zero for large values of $\hat{F}$. Therefore, we implement eq. (5.17) iteratively to normalise in a distributed way. In iteration $f \in \{1, \cdots, \hat{F}\}$, we calculate

$$p_{\boldsymbol{h}_f|M}(\boldsymbol{b}|m) = p^{(f-1)/f}_{\boldsymbol{h}_{f-1}|M}(\boldsymbol{b}|m) \begin{cases} p^{1/f}_{h^{(f)}|M}(b|m) & \text{After detection gap} \\ p^{1/f}_{h^{(f)}|h^{(f-1)},M}(b|b_0, m) & \text{Elsewhere} \end{cases}, \tag{5.18}$$

where $\boldsymbol{h}_f$ is the subvector of $\boldsymbol{h}$ on frames 1 to f, and $\boldsymbol{h}_0$ can be set to any finite value. The final result $\boldsymbol{h}_{\hat{F}} = \boldsymbol{h}$ provides the solution of eq. (5.17), normalised and corrected for missing detections.

5.1.6. MIXTURE MODELS

The assumption was stated in section 5.1.5 that each Markov model belongs to a single workflow phase. The reverse we do not assume, i.e., multiple Markov models can indicate the same phase. A workflow phase $A \in \mathcal{A}$ from a predefined set $\mathcal{A}$ can be represented by a mixture model. The mixture model consists a set of weighted Markov models $\mathcal{M}_A$ with weights

$$p_{M|A}(m|a). \tag{5.19}$$

Given A, we calculate the probability of a history vector $\boldsymbol{h}$ appearing as

$$p_{\boldsymbol{h}|A}(\boldsymbol{b}|a) = \sum_{m \in \mathcal{M}_a} p_{\boldsymbol{h},M|A}(\boldsymbol{b}, m|a). \tag{5.20}$$

As $\boldsymbol{h}$ is fully generated by a Markov model, which in turn is defined fully by A, we can write

$$p_{\boldsymbol{h},M|A}(\boldsymbol{b}, m|a) = p_{\boldsymbol{h}|M}(\boldsymbol{b}|m)p_{M|A}(m|a), \tag{5.21}$$

which components are given by eqs. (5.18) and (5.19).

5.1.7. WORKFLOW PHASE TRANSITIONS

In practice, a workflow phase is not often recognisable from an arbitrary window. It is necessary to build memory into the algorithm that extends beyond the window length F. Workflow dynamics change throughout a procedure. Workflow phases or phase transitions are more- or less likely depending on elapsed time. This knowledge should be used during phase classification. We use a time-dependent Markov model to achieve this, defining probabilities

$$p_{A_w}(a_w, w), \tag{5.22}$$

$$p_{A_w|A_{w-1}}(a_w, w|a_{w-1}), \tag{5.23}$$

where A_w is the workflow phase shown in window w. We evaluate a version of the Viterbi algorithm [24]:

$$p_{A_w,\boldsymbol{h}_w}(a_w, \boldsymbol{b}_w) \propto \sum_{a_{w-1} \in \mathcal{A}} p_{h_w|A_w}(b_w|a_w)p_{A_w|A_{w-1}}(a_w, w|a_{w-1}) \times p_{A_{w-1},\boldsymbol{h}_{w-1}}(a_{w-1}, \boldsymbol{b}_{w-1}), \tag{5.24}$$

where h_w is the set of history vectors in window w, $\boldsymbol{h}_w$ is the set of history vectors in windows before w, prior $p_{h_w}(b_w)$ is assumed to be uniform, and h_w is assumed to be independent of A_{w-1}. the first two components are given in eqs. (5.20) and (5.23), and the last is obtained iteratively. For the first iteration we write

$$p_{A_1,h_1}(a_1, b_1) = p_{h_1|A_1}(b_1|a_1)p_{A_1}(a_1, 1), \tag{5.25}$$

which components are given by eqs. (5.20) and (5.22).

Every iteration, eq. (5.24) or eq. (5.25) gives the a-posteriori probability of the newest window being in a certain phase. This probability is maximised over $\mathcal{A}$ to obtain a classification.

5.1.8. TRAINING

Inference requires a set of Markov models, grouped into mixture models. Each mixture model is trained on the labelled windows from section 5.1.3 using Expectation Maximisation [25]. First, the probabilities from eqs. (5.13) and (5.14) are initialised randomly per Markov model. The window order is randomised during training. Two steps are repeated until no more windows are left in the training set:

1. Expectation: From a new batch of windows $\mathcal{W}_A$ belonging to phase A, each tracklet is scored per Markov model $M \in \mathcal{M}_A$ using eq. (5.18). Maximising the obtained probabilities over $\mathcal{M}_A$ forms a cluster of tracklets per Markov model. Tracklets with fewer than F_{min} samples are discarded.

2. Maximisation: A new Markov model is extracted empirically from each cluster by counting the number of bin appearances- and transitions. In the case that a cluster contains fewer than T_{min} tracklets, the original Markov model is kept instead. This yields a new set of models $\mathcal{M}_A$.

This process can be repeated for a number of epochs, re-adding all windows to the training set in random order each time. Note that Expectation Maximisation only makes sense when $|\mathcal{W}_A| \gg |\mathcal{M}_A|$, since each cluster should have at least one tracklet to extract empirical data. After the last iteration, counting the number of tracklets per cluster yields empirical Markov model weights for eq. (5.19).

The time-dependent phase transition Markov model in eqs. (5.22) and (5.23) was estimated empirically by counting the annotated phases per window. To obtain a smooth result, a long window length $F_{\text{prior}} > F$ was used. Due to scarce workflow phase transitions, this approach yielded very low transition probabilities. This makes it unlikely for the algorithm to change its prediction between windows. To fix this, transition probabilities were scaled to have a fixed standard deviation σ_t. Finally, the probabilities were smoothed over time using a Savitzky-Golay filter [26] to reduce noise.

5.1.9. RE-ADDING POSES AND KEYPOINTS

Variables $K \in \mathcal{K}$, $p \in \mathcal{P}$ and r were assumed fixed to simplify notation. By training and applying the method separately for each keypoint

class K, different models are obtained that estimate workflow phase independently. This has two advantages:

- Each keypoint class moves differently, providing a unique source of workflow information.
- Separation clarifies the contribution per keypoint, contrasting the 'black-box' nature of e.g. a Neural Network.

r can be varied in the same way, extracting different aspects of motion and providing transparancy on their effectiveness. The same should not be done for p, as tracking IDs are not guaranteed to be accurate over long timespans and change between videos. Predictions of different models are averaged as an ensemble to obtain a final estimate.

5.1.10. EXPERIMENTS

Table 5.2.: Hyperparameters of three models, each trained on another history vector feature. The number of bins for inter-distance history vectors were limited by system memory.

Parameter	Symbol		Value	
Window step	S		2 s	
Window length	F		10 s	
Prior window length	F_{prior}		30 s	
Prior deviation	σ_t		0.15	
Mixture components	$\lvert\mathcal{M}_A\rvert$		50	
Batch size	$\lvert\mathcal{W}_A\rvert$		500	
Training epochs	–		10	
Min. tracklet samples	F_{min}		5	
Min. model tracklets	T_{min}		4	
History reference	r	Velocity	Intra-distance	Inter-distance
Magnitude bins	B_ρ	5	10	5
Orientation bins	B_θ	8	16	8
First magnitude bin	ρ_{first}	0.5	0.5	0.5
Last magnitude bin	ρ_{last}	15	200	200
Frame step	Δf	1	–	–

We train three models to recognise phases $\mathbb{A}$, $\mathbb{B}$, $\mathbb{C}$ and $\mathbb{D}$: one for each type of history vector. The used hyperparameters are shown in

table 5.2. The models are trained and tested on the seven videos from section 5.1.1. Testing is done both excluding and including the phase and transition priors from section 5.1.7.

Results over the whole dataset are summarised in confusion matrices. We construct each matrix by adding the predicted phase probabilities of each window. For example, if the four phases were predicted with probabilities 0.5, 0.2, 0.21 and 0.09, then these exact values are added to the corresponding confusion matrix row. Afterwards, each row is normalised to sum to one. Additionally, we report accuracy after making a hard decision on workflow phase, dividing the number of true predictions by the number of windows. True- and false predictions are counted using the maximum predicted probabilities. Accuracy from the velocity- and inter-distance models are reported separately per keypoint, and from the intra-distance model per keypoint pair. Results of keypoints that come in pairs, e.g., the two eyes, are averaged. Results of the velocity- and intra-distance models are averaged over poses, and for the inter-distance model over pose pairs.

Phase predictions of the velocity- and inter-distance models are shown over time for one example video. These are again presented per keypoint, where results of keypoints that come in pairs are averaged. In these figures, the ground truth timeline is shown for reference.

5.2. RESULTS

This section shows the results of the experiments of section 5.1.10. First, section 5.2.1 presents the normalised confusion matrices per tested model. Section 5.2.2 continues with the accuracy of each model, presented per keypoint. Finally, section 5.2.3 shows the phase predictions per keypoint during one example procedure.

5.2.1. CONFUSION MATRICES

Confusion matrices for all tested models are shown in table 5.3. When not imposing the time-dependent prior on workflow phases and transitions, the conditioned predictions approach a uniform distribution. When imposing the prior, a slight prediction bias is observed towards phases $\mathbb{A}$ and/or $\mathbb{C}$ when using intra- and inter-distance velocity vectors, regardless of the ground truth.

5.2.2. ACCURACY

The accuracy of the velocity- and inter-distance models is presented in table 5.4. The metric is shown per keypoint, and for the model overall. Without imposing the time-dependent workflow phase prior, 0.36 accuracy is measured using velocity history for all keypoints. The

Table 5.3.: Confusion matrices of the six tested models. Phase probabilities were summed over all windows, and normalised such that their rows sum to one.

(a) Velocity, excluding priors

Ground truth	𝔸	𝔹	ℂ	𝔻
𝔸	.25	.25	.25	.25
𝔹	.25	.25	.25	.25
ℂ	.25	.25	.25	.25
𝔻	.25	.25	.25	.25

(b) Intra-distance, excluding priors

Ground truth	𝔸	𝔹	ℂ	𝔻
𝔸	.25	.25	.25	.25
𝔹	.25	.25	.25	.25
ℂ	.25	.25	.25	.25
𝔻	.25	.25	.25	.25

(c) Inter-distance, excluding priors

Ground truth	𝔸	𝔹	ℂ	𝔻
𝔸	.25	.25	.25	.25
𝔹	.25	.25	.25	.25
ℂ	.25	.25	.25	.25
𝔻	.25	.25	.25	.25

(d) Velocity, including priors

Ground truth	𝔸	𝔹	ℂ	𝔻
𝔸	.25	.25	.25	.25
𝔹	.25	.25	.25	.25
ℂ	.25	.25	.25	.25
𝔻	.25	.25	.25	.25

(e) Intra-distance, including priors

Ground truth	𝔸	𝔹	ℂ	𝔻
𝔸	.25	.25	.26	.24
𝔹	.26	.25	.26	.24
ℂ	.26	.25	.26	.24
𝔻	.25	.25	.26	.24

(f) Inter-distance, including priors

Ground truth	𝔸	𝔹	ℂ	𝔻
𝔸	.26	.25	.25	.24
𝔹	.27	.25	.25	.23
ℂ	.27	.25	.25	.23
𝔻	.26	.25	.25	.24

Table 5.4.: Accuracy of the velocity- and inter-distance-based models per keypoint. The models were run with- and without imposing phase and transition priors. Results are averaged over keypoints that come in pairs, e.g., the two eyes. For the velocity model results are averaged over poses, and for the inter-distance model over pose pairs.

	Excluding priors		Including priors	
Keypoint	Velocity	Inter-distance	Velocity	Inter-distance
Nose	.36	.28	.45	.32
Eyes	.36	.30	.45	.32
Ears	.36	.33	.46	.32
Shoulders	.36	.29	.46	.32
Elbows	.36	.32	.44	.34
Wrists	.36	.33	.44	.33
Hips	.36	.31	.42	.32
Knees	.36	.34	.41	.32
Ankles	.36	.38	.35	.32
Total	.35	.34	.39	.32

total model reports an accuracy of 0.35. The inter-distance model reports varying accuracies per keypoint. Here, the nose and shoulders yield the lowest accuracy of up to 0.29, and the knees and ankles the highest above 0.34. The overall inter-distance model scores 0.34 accuracy.

When imposing priors, the velocity model now shows scores from 0.35 for the ankles to 0.46 for the ears and shoulders. This model scores 0.39 accuracy: a 4 percentage point increase with respect to the prior-less model. Most keypoints in the inter-distance model now yield 0.32 accuracy. The only exceptions are the elbows and wrists, which score 0.34 and 0.33 respectively. The whole model scores 0.32 accuracy, which is 2 percentage point lower than before imposing the priors.

Accuracy for the intra-distance model is shown in table 5.5. Scores are reported per keypoint pair. Before imposing priors, there is little variation between pairs, each yielding between 0.31 and 0.33 accuracy. This total model reports 0.31 accuracy, which is below the prior-less velocity- and inter-distance models. Imposing priors increases variation. Now, pairs including the ankles yield the lowest scores: from 0.28 to 0.30. The pairs between the shoulders and nose yield the best score of 0.42. All remaining pairs score between 0.34 and 0.39. The overall

Table 5.5.: Accuracy of the intra-distance-based model per keypoint pair. The models were run with- and without imposing phase and transition priors. Results are averaged over poses, and keypoints that come in pairs, e.g., the two eyes. The accuracies of the ensemblees over all keypoint pairs are 0.31 and 0.38 excluding and including the priors, respectively.

		Eyes	Ears	Shoulders	Elbows	Wrists	Hips	Knees	Ankles
Excluding priors	Nose	.33	.32	.33	.32	.32	.32	.32	.33
	Eyes		.32	.32	.32	.31	.32	.31	.32
	Ears			.33	.31	.31	.32	.32	.32
	Shoulders				.32	.32	.32	.32	.33
	Elbows					.32	.32	.33	.33
	Wrists						.33	.32	.32
	Hips							.32	.33
	Knees								.32
Including priors	Nose	.38	.39	.42	.37	.37	.36	.38	.30
	Eyes		.36	.38	.37	.36	.39	.34	.29
	Ears			.39	.34	.35	.38	.36	.28
	Shoulders				.35	.36	.35	.35	.28
	Elbows					.36	.38	.38	.30
	Wrists						.39	.36	.28
	Hips							.36	.30
	Knees								.28

model yields 0.38 accuracy: 1 percentage point below the velocity model, and 6 percentage point better than inter-distance.

5.2.3. QUALITATIVE RESULT

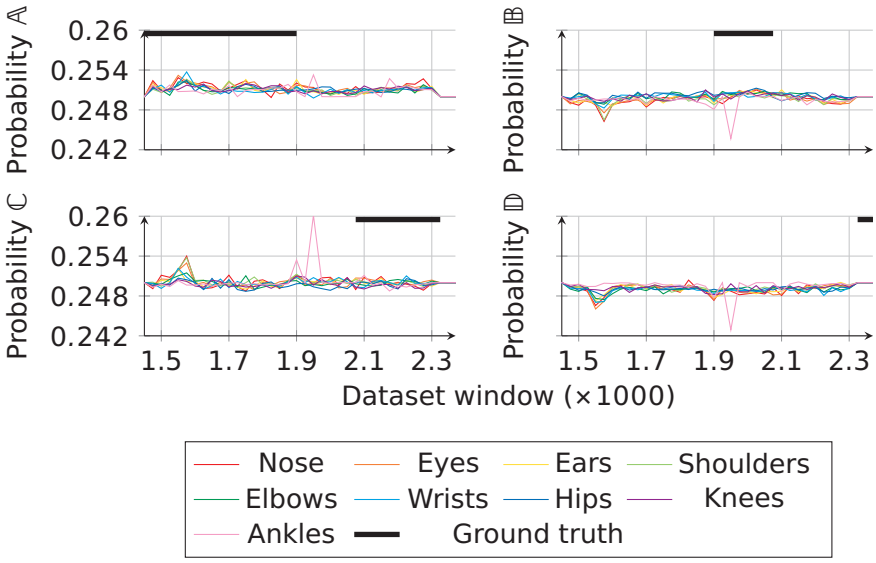


Figure 5.4.: Predicted phase probabilities over time using velocity history vectors including priors, during one example procedure. Results are presented per keypoint class. Annotated periods are shown per phase.

Predicted phase probabilities and the ground truth are shown during an example procedure in figs. 5.4 and 5.5. These figures show velocity- and inter-distance-based predictions, respectively. The velocity models predict phase probabilities between 0.242 to 0.26. Around the 1950th window, a spike can be seen where the ankles predict phase ℂ, although phase 𝔹 was annotated. The inter-distance models show more variation, especially in the prediction of phases 𝔸 and 𝔻. Both models show a slight bias towards phase 𝔸. No clear relation between the ground truth and predictions seems present in the figures.

5.3. DISCUSSION

In this work, we investigated workflow phase detection from human pose history vectors. For interpretability, each feature produced a separate prediction, after which a final conclusion was established using an ensemble.

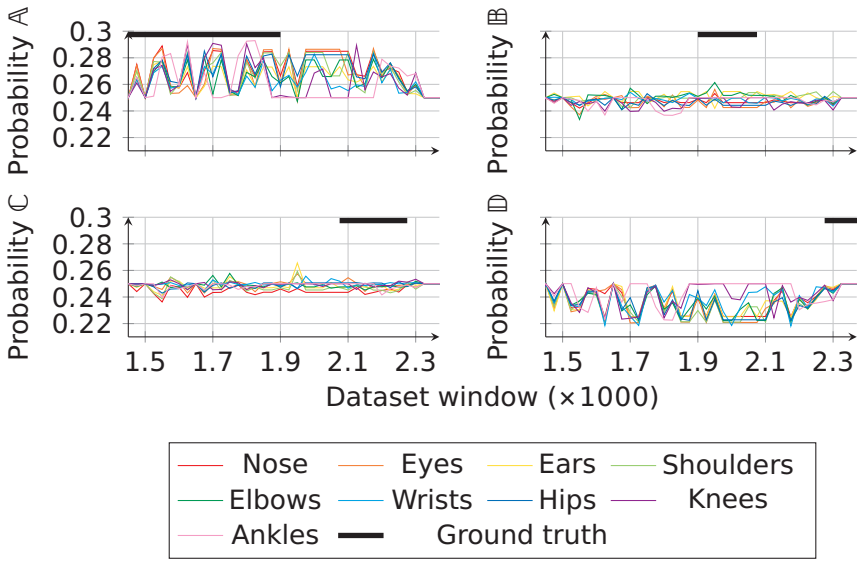


Figure 5.5.: Predicted phase probabilities over time using inter-distance history vectors including priors, during one example procedure. Results are presented per keypoint class. Annotated periods are shown per phase.

The confusion matrices and qualitative example revealed that roughly uniform phase probabilities were predicted by every model. Each separate keypoint (pair) yielded approximately uniform predictions, which explains the uniformity of their ensembles. No clear correlation was observed between the ground truth and predictions. There was a slight difference between models, where the inter-distance model outputs showed larger deviations than those of the velocity models.

Although phase predictions were approximately uniform, the reported accuracy was always higher than 0.25. This is caused by the slight bias of all models towards phase $\mathbb{A}$, often causing them to classify this phase over the others. As the majority of windows was annotated as phase $\mathbb{A}$, always predicting this phase yields a better accuracy than random guessing. Adding a time-dependent phase prior introduced bias towards phase $\mathbb{C}$ in some models. The prior may enable models to make a more educated guess based on elapsed time, increasing accuracy slightly.

Differences in accuracy were present between history vector types, keypoints, and keypoint pairs. Considering the near-uniform predicted phase probabilities, the accuracy-based reasoning in this paragraph should be interpreted as speculative. Before imposing phase priors, the intra-distance model yielded the worst accuracy. This suggests that relations between keypoints of the same class might be contain more

workflow information than those between keypoints of different classes. In the inter-distance model, the ankles yielded the highest accuracy. Ankles are a relatively robust indicator of positioning, as posture affects other keypoint locations. The ankles of the cardiologist and scrub nurse were not visible from the used viewpoint during phases $\mathbb{B}$ and $\mathbb{C}$, but this could have been compensated by them not walking during these phases. When imposing priors, inter-distance vectors yielded the worst results. This suggests that time information adds workflow context to keypoint relations within an individual, but not to those between individuals. This effect worsened the accuracy of ankle keypoints, which may have served the same purpose of estimating procedure progression by looking at personnel locations.

A flaw of the proposed algorithm is that, to achieve interpretability, it barely associates information between different keypoints or pairs. Looking at a single keypoint (pair) at a time did not provide sufficient information to recognise a workflow phase. Whereas the original velocity history vectors predicted activities based on many keypoints, the used human pose tracklets may have been too sparse. Although history vectors could still prove to be a good input feature, either 1) more keypoints should be followed using e.g. SuperGlue [27], and/or 2) a more associative classifier should be used. Neural networks excel at association, but their decision-making process is poorly interpretable. References [15, 28, 29] successfully used long short-term memory- and generative adversarial networks to detect instrument passing, staff attention, and workflow anomalies from human poses in the operating room. As a compromise, a neural network could be used in conjunction with interpretation methods [16]. Alternatively, a neural network design specifically for this problem could clarify the purpose of each neuron. Graph neural networks could be a natural choice for modelling relations between keypoints and/or individuals [11, 12]. Other information sources, e.g., device logs could be added to create a multimodal learning algorithm. The quality of poses could be improved by implementing multi-view information, or the bin distribution of history vectors could be tweaked with a parameter search.

Besides procedure duration, more prior knowledge could be employed. For instance, knowledge of the roles of individuals could play a large role. The patient is walking during phases $\mathbb{A}$ and $\mathbb{D}$, and lying down during $\mathbb{B}$ and $\mathbb{C}$. After making a distinction based on such prior knowledge, more refined phases or activities could be extracted using e.g. motion of the cardiologist. As workflows in many medical procedures are highly standardised, such prior knowledge should be readily available.

The proposed method is too unreliable to base any workflow support systems on it. However, it demonstrates the added value of an interpretable method. The importance of procedure duration as a feature is apparent from the results. The varying contributions of

different keypoints and history vectors can inform the design of future systems. For instance, when looking at individual keypoints, i.e., velocity history in combination with procedure duration, upper body keypoints seem to provide most information. When the procedure duration is not available, ankle accuracy suggest that the historical distance between individuals might present an alternative feature. Accuracy from intra-distance history vectors suggests that keypoints that lie close do not necessarily have the most descriptive relations. For example, in combination with procedure duration, the elbows appeared to have a more descriptive relation with the knees than the wrists. Therefore, when designing a pose-based phase classifier, counterintuitive relations between keypoints and individuals should be considered.

Future work can consider recommendations from the previous paragraph in the development of monitoring-based workflow phase detection. Monitored phases could be presented to staff in personalised dashboards. For example, timelines like those in figs. 5.4 and 5.5, or statistics on time spent per phase could be presented. A centralised system could analyse personnel motion and workflow information from multiple teams and hospitals to identify best practices. After a procedure, staff could automatically receive personalised feedback based on their performance, and suggestions to improve their workflow. Automation enables an anonymous implementation, where videos are never viewed by humans and workflow metrics are shown only to the considered staff member.

In addition to feedback systems, workflow monitoring can provide context-awareness for automated real-time support. Information shown on the monitor could be adapted to the current workflow step, or device settings such as C-arm position could be suggested and implemented after approval by the cardiologist. Personalised feedback, and effective cooperation between man and machine, can streamline process efficiency and safety in the Cath Lab without increasing the workload. Implemented effectively, this will have a positive effect on hospital management, staff work enjoyment, and patient care.

5.4. CONCLUSION

The tested method did not have the capacity to extract useful workflow information. A method is needed that associates between features more effectively. Results demonstrate an importance of feature selection, where procedure duration is vital and different human keypoint relations should be considered. Detecting not only a workflow phase, but the underlying staff dynamics, is key to applications in feedback- and support systems.

REFERENCES

[1] R. Marjamaa, A. Vakkuri, and O. Kirvelä. “Operating Room Management: Why, How and by Whom?” In: *Acta Anaesthesiol. Scand.* 52.5 (Apr. 2008), pp. 596–600. doi: 10.1111/j.1399-6576.2008.01618.x.

[2] C. B. E. Halbeis and A. Schubert. “Staffing the Operating Room Suite: Perspectives from Europe and North America on the Role of Different Anesthesia Personnel”. In: *Anesthesiol. Clinic.* 26.4 (Dec. 2008), pp. 637–663. doi: 10.1016/j.anclin.2008.07.002.

[3] A. M. Schouten, S. M. Flipse, K. E. van Nieuwenhuizen, F. W. Jansen, A. C. van der Eijk, and J. J. van den Dobbelsteen. “Operating Room Performance Optimization Metrics: a Systematic Review”. In: *J. Med. Syst.* 47.1 (Dec. 2023), p. 19. doi: 10.1007/s10916-023-01912-9.

[4] A. J. Fong, M. Smith, and A. Langerman. “Efficiency Improvement in the Operating Room”. In: *J. Surg. Res.* 204.2 (Aug. 2016), pp. 371–383. doi: 10.1016/j.jss.2016.04.054.

[5] C. von Schudnat, K.-P. Schoeneberg, J. Albors-Garrigos, B. Lahmann, and M. De-Miguel-Molina. “The Economic Impact of Standardization and Digitalization in the Operating Room: A Systematic Literature Review”. In: *J. Med. Syst.* 47 (Dec. 2023), p. 55. doi: 10.1007/s10916-023-01945-0.

[6] K. N. Timoh, A. Huaulme, K. Cleary, M. A. Zaheer, V. Lavoué, D. Donoho, and P. Jannin. “A Systematic Review of Annotation for Surgical Process Model Analysis in Minimally Invasive Surgery based on Video”. In: *Surg. Endosc.* 37 (May 2023), pp. 4298–4314. doi: 10.1007/s00464-023-10041-w.

[7] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, M. Hashizume, D. Katic, H. Kenngott, M. Kranzfelder, A. Malpani, K. März, T. Neumuth, N. Padoy, C. Pugh, N. Schoch, D. Stoyanov, R. Taylor, M. Wagner, G. D. Hager, and P. Jannin. “Surgical Data Science for Next-generation Interventions”. In: *Nat. Biomed. Eng.* 1 (Sept. 2017), pp. 691–696. doi: 10.1038/s41551-017-0132-7.

[8] C. R. Garrow, K.-F. Kowalewski, L. Li, M. Wagner, M. W. Schmidt, S. Engelhardt, D. A. Hashimoto, H. G. Kenngott, S. Bodenstedt, S. Speidel, B. P. Müller-Stich, and F. Nickel. “Machine Learning for Surgical Phase Recognition: A Systematic Review”. In: *Annal. Surg.* 273.4 (Apr. 2021), pp. 684–693. doi: 10.1097/SLA.0000000000004425.

[9] Mayo Clinic Staff. *Coronary Angiogram*. Dec. 2023. url: https://www.mayoclinic.org/tests-procedures/coronary-angiogram/about/pac-20384904.

[10] G. Saleem, U. I. Bajwa, and R. H. Raza. “Toward Human Activity Recognition: a Survey”. In: *Neural Comput. Appl.* 35 (Feb. 2023), pp. 4145–4182. doi: 10.1007/s00521-022-07937-4.

[11] L.-F. Wu, Q. Wang, M. Jian, Y. Qiao, and B.-X. Zhao. “A Comprehensive Review of Group Activity Recognition in Videos”. In: *Int. J. Autom. Comput.* 18 (June 2021), pp. 334–350. doi: 10.1007/s11633-020-1258-8.

[12] H.-C. Nguyen, T.-H. Nguyen, and V.-H. Scherer Rafałand Le. “Deep Learning for Human Activity Recognition on 3D Human Skeleton: Survey and Comparative Study”. In: *Sens.* 23.11 (May 2023), p. 5121. doi: 10.3390/s23115121.

[13] C. Wang and J. Yan. “A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition”. In: *IEEE Access* 11 (June 2023), pp. 53880–53898. doi: 10.1109/ACCESS.2023.3282311.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. “Microsoft COCO: Common Objects in Context”. In: *Eur. Conf. Comput. Vis.* Springer, Sept. 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[15] K. Yokoyama, G. Yamamoto, C. Liu, K. Kishimoto, and T. Kuroda. “Operating Room Surveillance Video Analysis for Group Activity Recognition”. In: *Adv. Biomedic. Eng.* 12 (2023), pp. 171–181. doi: 10.14326/abe.12.171.

[16] A. Abusitta, M. Q. Li, and B. C. M. Fung. “Survey on Explainable AI: Techniques, Challenges and Open Issues”. In: *Expert Syst. Appl.* 255.C (Dec. 2024), p. 124710. doi: 10.1016/j.eswa.2024.124710.

[17] R. Messing, C. Pal, and H. Kautz. “Activity Recognition using the Velocity Histories of Tracked Keypoints”. In: *IEEE Int. Conf. Comput. Vis.* IEEE, Sept. 2009, pp. 104–111. doi: 10.1109/ICCV.2009.5459154.

[18] Noldus. *The Observer XT*. url: https://www.noldus.com/observer-xt.

[19] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu. “AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 45.6 (June 2023), pp. 7157–7173. doi: 10.1109/TPAMI.2022.3222784.

[20] J. Redmon and A. Farhadi. *YOLOv3: An Incremental Improvement*. Apr. 2018. doi: 10.48550/arXiv.1804.02767. arXiv: 1804.02767v1 [cs.CV].

[21] K. He, X. Zhang, S. Ren, and J. Sun. “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 37.9 (Sept. 2015), pp. 1904–1916. doi: 10.1109/TPAMI.2015.2389824.

[22] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *Proc. 29th IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[23] R. M. Butler, T. S. Vijfvinkel, E. Frassini, S. van Riel, C. Bachvarov, J. Constandse, M. van der Elst, J. J. van den Dobbelsteen, and B. H. W. Hendriks. “2D Human Pose Tracking in the Cardiac Catheterisation Laboratory with BYTE”. In: *Med. Eng. Phys.* 135 (Jan. 2025), p. 104270. doi: 10.1016/j.medengphy.2024.104270.

[24] G. D. Forney. “The Viterbi Algorithm”. In: *Proc. IEEE* 61.3 (Mar. 1973), pp. 268–278. doi: 10.1109/PROC.1973.9030.

[25] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm”. In: *J. R. Stat. Soc., Ser. B (Stat. Methodolog.)* 39.1 (1977), pp. 1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x.

[26] A. Savitzky and M. J. E. Golay. “Smoothing and Differentiation of Data by Simplified Least Squares Procedures”. In: *Anal. Chem.* 36.8 (July 1964), pp. 1627–1639. doi: 10.1021/ac60214a047.

[27] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. “SuperGlue: Learning Feature Matching with Graph Neural Networks”. In: *Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2020, pp. 4937–4946. doi: 10.1109/CVPR42600.2020.00499.

[28] K. Yokoyama, G. Yamamoto, C. Liu, O. Sugiyama, L. H. Santos, and T. Kuroda. “Recognition of Instrument Passing and Group Attention for Understanding Intraoperative State of Surgical Team”. In: *Adv. Biomedic. Eng.* 11 (2022), pp. 37–47. doi: 10.14326/abe.11.37.

[29] K. Yokoyama, G. Yamamoto, C. Liu, K. Kishimoto, Y. Mori, and T. Kuroda. “Individual Activity Anomaly Estimation in Operating Rooms Based on Time-Sequential Prediction”. In: *World Congr. Med. Health Inform.* IOS Press, July 2023, pp. 284–288. doi: 10.3233/SHTI230972.


6

QUANTIFYING INTERACTION WITH THE OPERATING TABLE

Perioperative staff shortages are a problem in hospitals worldwide. Keeping the staff content and motivated is a challenge in the busy hospital setting of today. New operating room technologies aim to increase safety and efficiency. This causes a shift from interaction with patients to interaction with technology. Objectively measuring this shift could aid the design of supportive technological products, or optimal planning for high-tech procedures. 35 gynaecological procedures of three different technology levels are recorded: open- (OS), minimally invasive- (MIS) and robot-assisted (RAS) surgery. We annotate interaction between staff and the patient. An algorithm is proposed that detects interaction with the operating table from staff posture and movement. Interaction is expressed as a percentage of total working time. The proposed algorithm measures operating table interactions of 70.4 %, 70.3 % and 30.1 % during OS, MIS and RAS. Annotations yield patient interaction percentages of 37.6 %, 38.3 % and 24.6 %. Algorithm measurements over time show operating table- and patient interaction peaks at anomalous events or workflow phase transitions. The annotations show less operating table- and patient interaction during RAS than OS and MIS. Annotated patient interaction and measured operating table interaction show similar differences between procedures and workflow phases. The visual complexity of operating rooms complicates pose tracking, deteriorating the algorithm input quality. The proposed algorithm shows promise as a component in context-aware event- or workflow phase detection.

Parts of this chapter were published in Int. J. Comput. Assist. Radiol. Surg. **vol** (2025),

Technology plays an increasingly large role in the Operating Room (OR) [2]. New technologies aim to improve patient safety and procedure efficiency [3]. The adoption of robot-assisted surgery (RAS) has grown in the last few decades. Currently, RAS requires larger teams and more time to perform than minimally invasive surgery (MIS) or open surgery (OS) [4].

RAS, MIS and OS demand different skillsets from surgical staff [5]. Procedures of technical nature shift the focus from direct patient care towards technical activities [6, 7]. This shift impacts work perception- and satisfaction of the staff [8].

Added complexity and a shift away from care add stress to an already stressful environment [6–8]. This can diminish quality of care and staff wellbeing. Consequences like communication difficulties, feelings of isolation, and anxiety are quoted. Each of these contributes negatively to patient safety.

Shortages of perioperative staff and high turnover rates are a worldwide concern [8]. Literature identifies workload as a major cause [9]. Beside addressing workload, workflow insights can lead to effective staff deployment and streamlined processes [3].

New technologies should ideally support healthcare professionals without getting in the way or inducing stress. If a technology causes severe changes in workflow, or increases procedure complexity, its design might leave room for improvement. Specifically, some technologies may demand much attention from personnel, thereby shifting focus from direct patient care towards technical tasks.

Knowledge about the effects of technologies on perioperative workflow can aid in the design of new products and support systems. To map these effects, an interesting metric is the time spent on direct patient care. One possible approach to measuring this metric is automatic monitoring of personnel activities in procedure videos. Such monitoring could be deployed in hospitals on a large scale. Outcomes could yield relations between procedure technology levels and perceived workload.

Insights obtained by monitoring from many hospitals could help in the design of future OR technologies. For example, if much time is consistently spent configuring a device during procedures, this reveals an opportunity where user-friendliness can be improved. A new iteration of the product could e.g. carry out the configuration autonomously, or simplify it by making suggestions on its own. This way, the technology assumes a more supportive role, without requiring much attention from the staff. Another application is to optimise planning and logistics for e.g. turnover time and staff wellbeing [10]. Device placement could be updated for better ergonomics or workflow efficiency. Tasks could be divided differently to distribute workload more uniformly over the surgical team.

Computer vision for automated OR monitoring is an upcoming research

topic [11]. Bounding box- or pose detection can localise individuals in video. Pose trackers infer bodypart—or keypoint—coordinates from all persons in a video on every frame. Detection confidence is scored per keypoint, and each individual is assigned a unique identifier (ID) for re-identification between frames. Most state-of-the-art 2D pose trackers rely on neural networks that need training on annotated images. Some authors provide models that were pretrained on datasets like COCO [12] or MPII [13]. Important to consider is that monitoring itself could introduce discomfort or stress for OR staff. Monitoring systems should be designed carefully and non-intrusively, in a way that does not hinder personnel comfort and wellbeing.

The OR shows visual differences from general-purpose datasets. Clutter, occlusion and visually similar clothing complicate detection- and tracking. It cannot be assumed that algorithms trained on general situations perform well in the OR. Reference [14] presents an annotated dataset with recordings of real surgeries. To our knowledge, this is the only such public dataset at the time of writing.

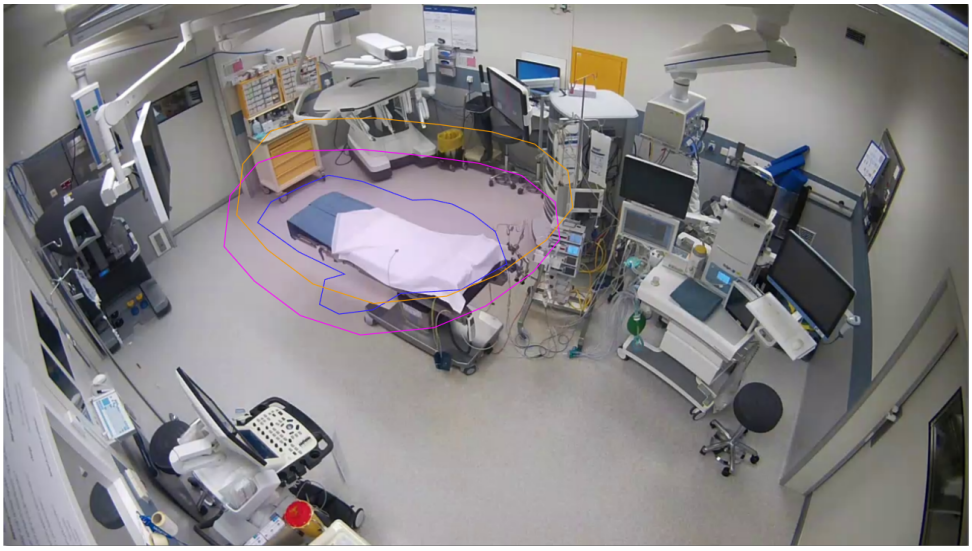
This work presents a first exploration to quantify interaction between staff and the patient from monitoring videos, during procedures of varying technology levels. We take a multimodal approach, where a computer vision algorithm and manual annotations provide complementing measurements of interaction with the operating table and patient. To our knowledge, no automated monitoring tool that measures such perioperative interaction exists in literature at the time of writing. Patient interaction is annotated based on observed intent, and human pose tracklet motion and position are constrained to automatically classify operating table interaction. An interaction metric is designed specifically to counteract bias from missing pose detections.

Section 6.1 describes our dataset, classification of operating table interaction, and experiments. The outcomes are shared in section 6.2 and discussed in section 6.3. Finally, section 6.4 presents our conclusions.

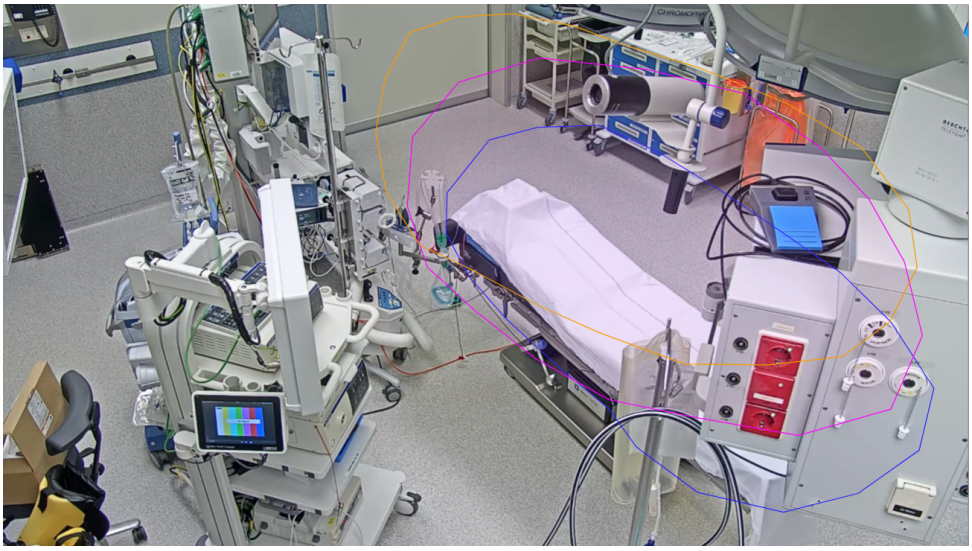
6.1. METHODS

6.1.1. DATASET

Videos were recorded in two LUMC ORs during 35 OS, MIS or RAS gynaecological procedures, from the viewpoints shown in fig. 6.1. The study was approved by a local medical ethics committee, and all included patients gave informed consent. MIS- and OS procedures were filmed using the same synchronised four-camera setup with a resolution of 1920 px $\times$ 1080 px per viewpoint. RAS—carried out with the da Vinci surgical platform—was filmed with two synchronised cameras with a larger field of view and a resolution of 1280 px $\times$ 720 px.



(a) RAS surgery viewpoint



(b) MIS and OS surgery viewpoint

Figure 6.1.: Recording viewpoints in the two ORs. Annotated regions are shown where the wrists (blue), shoulders (purple) and head (orange) must be for a person to interact with the operating table.

Table 6.1.: Annotated personnel actions, and their classification as interaction with the patient.

Annotated action	Label
Active at table	Patient interaction
Transferring instruments	
Wrapping DV system	
Active elsewhere	No patient interaction
Inactive	
Unpacking instruments	
Moving cart	
Absent	Absence
Action unknown	

Table 6.2.: Annotated workflow phases.

Phase	Description
Induction	Anaesthetic administration
Preparation	Surgical preparations
Surgery	The intervention
Recovery	Waking before departure

Each recording was started at anaesthetic induction, and ended after recovery in the OR. In each OR, the camera with the clearest view of the operating table area was selected for interaction quantification. The resulting viewpoints are shown in fig. 6.1.

In each procedure, areas were annotated where the wrists, shoulders and head of a person should be present for them to interact with the operating table. Two example annotations are shown in fig. 6.1. The wrists area was drawn loosely around the patient in a lying position. Shoulders and head areas were included to correct for the camera 2D projection of 3D scenes.

The personnel activities from the left side of table 6.1 were annotated for each person in the room. This was done by two annotators who were unaware of the automatic detection method under development. Annotated activies were grouped into the three categories on the right side of the table, for use in patient interaction classification. Finally, the workflow phases from table 6.2 were annotated to enable evaluations per phase.

6.1.2. POSE TRACKING

WWe use AlphaPose [15] to detect poses in the OR. AlphaPose applies a fast human bounding box detector [16, 17], after which features are extracted [18] and a Convolutional Neural Network (CNN) places a pose in each box. During training, a specific loss function and feature normalisation achieve keypoint translation- and scale invariance.

A tracker associates detected poses between video frames. AlphaPose includes an optional tracker that uses visual features. This strategy is unsuitable for the OR, as individuals here are dressed similarly. Instead, tracking is done with PoseBYTE [19], which uses only geometric information and prioritises confident detections. PoseBYTE adapts BYTE [20] to associate poses instead of bounding boxes using Object Keypoint Similarity (OKS) [12]. PoseBYTE discards tracklets that are not present for at least two subsequent frames. This compensates for the use of a low-threshold object detector by AlphaPose, which increases the risk of single-frame false positives.

Human bounding boxes are extracted from video with YOLOv3-SPP [16, 17], using features from ResNet152 [18]. A pose is detected in each bounding box using FastPose (DUC) [15], and tracked and refined using PoseBYTE [19]. The pose detector was pretrained by its authors on the COCO dataset [12], and we carried out no further training. PoseBYTE is no machine learning algorithm, and therefore requires no training.

6.1.3. DETECTING OPERATING TABLE INTERACTION

Our model for detecting personnel interaction with the operating table is visualised in table 6.3. When a person is standing still in the correct

Table 6.3.: Detecting interaction with the operating table from personnel position and movement.

		Position	
		By table	Elsewhere
Movement	Still	Other	Other
	Walking	Interaction	Other

position, this is assumed to signal interaction with the operating table. These two constraints are detailed in section 6.1.3.

MOVEMENT

To detect (lack of) movement, we calculate for each keypoint its displacement magnitude in px over a span of f_{motion} frames. Choosing a larger f_{motion} enables the capture of longer-term motion. To account for undetected keypoints, each pose is divided into subposes, for each of which movement is classified separately. A subpose s_{m} is defined to be still if at least a number $M^{(s_{\text{m}})}_{\text{keypoint}}$ of its keypoints $k^{(s_{\text{m}})}_{\text{m}} \subset s_{\text{m}}$ yields a displacement below a threshold $\tau^{(s_{\text{m}})}_{\text{m}}$. When a keypoint is not detected, or detected with a confidence below a threshold $\gamma^{(s_{\text{m}})}_{\text{m}}$, it is assumed to be still. A pose is defined to be still if at least a number M_{subpose} of its subposes are.

POSITION

We classify positioning using the annotated regions from section 6.1.1. Poses are divided into three subposes: i) the wrists, ii) the shoulders, and iii) the head—consisting of the nose, eyes and ears. Each subpose $s_{\text{p}} \in \{\text{wrists}, \text{shoulders}, \text{head}\}$ is classified to be by the table if at least a number $P^{(s_{\text{p}})}_{\text{keypoint}}$ of its keypoints $k^{(s_{\text{p}})}_{\text{p}} \subset s_{\text{p}}$ falls within the corresponding annotated region. Keypoints with a detection confidence below a threshold $\gamma^{(s_{\text{p}})}_{\text{p}}$ are not counted within any region. A pose is classified to be by the table if at least a number P_{subpose} of its subposes is.

6.1.4. ANNOTATING PATIENT INTERACTION

Detected operating table interaction is intended as a measure for patient interaction. However, it is not guaranteed that operating table interaction as defined in the algorithm indeed signals interaction with the patient. Therefore, the personnel activities from table 6.1 were annotated in the dataset in section 6.1.1. These annotations provide a separate measurement of actual interaction with the patient.

6.1.5. EXPERIMENTS

MODELS

Table 6.4.: Values for the parameters defined in section 6.1.3, used to detect pose movement based on two subposes.

Parameter	Value	
f_{motion}	5	
M_{subpose}	1	
s_{m}	1	2
$k_{\text{m}}^{(s_{\text{m}})}$	Shoulders	Head
$M_{\text{keypoint}}^{(s_{\text{m}})}$	1	1
$\tau_{\text{m}}^{(s_{\text{m}})}$	17.5 px	17.5 px
$\gamma_{\text{m}}^{(s_{\text{m}})}$	0.3	0.3

Table 6.5.: Parameters defined in section 6.1.3 used to detect pose position.

Parameter	Value		
P_{subpose}	2		
s_{p}	wrists	shoulders	head
$k_{\text{p}}^{(s_{\text{p}})}$	wrists	shoulders	nose, eyes, ears
$P_{\text{keypoint}}^{(s_{\text{p}})}$	1	1	2
$\gamma_{\text{p}}^{(s_{\text{p}})}$	0.3	0.3	0.15

The used algorithm parameters are shown in tables 6.4 and 6.5.

Constraining only a subset of subposes and keypoints compensates for undetected keypoints. Legs are excluded as they are detected least well. As arms can move during operating table interaction, their movement is not considered.

CLASSIFICATION

We measure the mean time fraction that personnel interacts with the operating table

$$F = \frac{1}{|P|} \sum_{p \in P} r(p), \tag{6.1}$$

$$r(p) = \begin{cases} 1 & \text{If } p \text{ interacts} \\ 0 & \text{Otherwise} \end{cases}, \tag{6.2}$$

where P is the set of all pose detections. Similarly, we measure the mean time fraction of movement by making $r(p)$ 1 when a pose is moving. The mean time fraction of patient interaction is measured using the annotations, where P includes all annotated activities not labelled as 'absence'.

This definition of F compensates for pose tracking errors in several ways. First of all, only detected poses contribute in eq. (6.1), i.e., false negatives do not affect F. Additionally, summing over all individuals removes any identity-specific information, mitigating re-identification errors. Finally, letting P cover a timespan—rather than a single frame—mitigates single-frame detection errors through time averaging.

Equation (6.1) introduces limitations as well. As detection accuracy varies between workflow phases, P will contain more accurate poses during some phases than others. Therefore, if P spans multiple workflow phases, this introduces a bias where some workflow phases affect F more than others. Another limitation is the equal treatment of all individuals in the room. Discarding information on person roles (e.g. surgeon, nurse, patient, spectator) means that all roles contribute equally to F. Patients and spectators therefore affect F, whereas our main interest is the interaction of only personnel with the table.

During experiments, we extract F for three selections of P per procedure type. First, we choose P to span all frames of all videos of the same procedure type jointly. The second experiment evaluates F per individual video, letting P span one video at a time. Finally, we evaluate the evolution of F over time within videos. Here, to maintain the time averaging effect, F is calculated over a sequence of time windows. Windows were chosen to have a length of 7500 f, with their start frames spaced 3750 f apart. Thus, two adjacent windows overlap with $7500\,\text{f} - 3750\,\text{f} = 3750\,\text{f}$.

Finally, we estimate pose detection performance by evaluating the quantity of detected human poses. The number of pose detections is

divided by the number of pose annotations per window. This should yield a value close to 1 if the numbers of annotated- and detected poses lie close. Individual pose detections cannot be verified without annotating their location. Note, therefore, that a value of 1 does not guarantee correct pose detections.

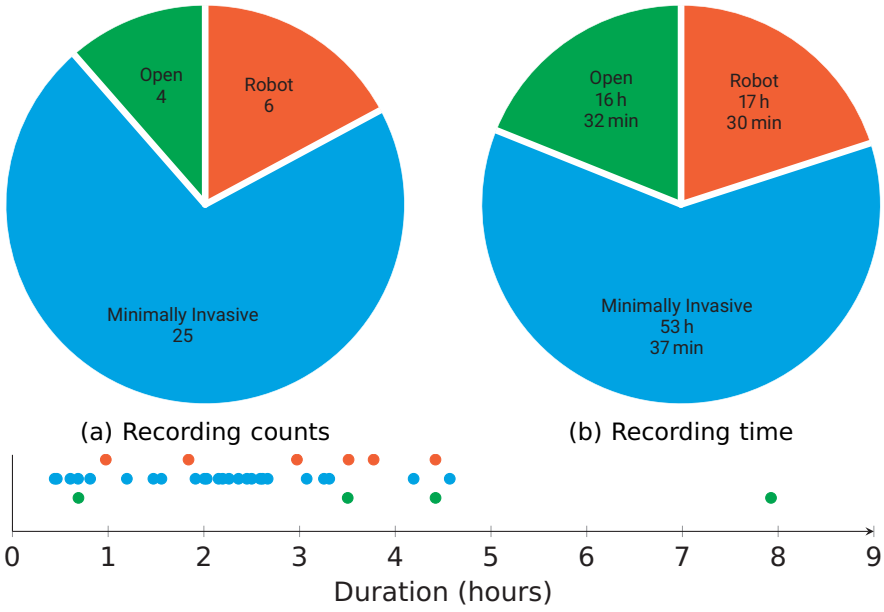
QUALITATIVE RESULTS

Qualitative results are shown with colour-coded pose detections. A pose is drawn green if the person was classified to interact with the operating table. If the person was in the right position, but moving too fast to interact, they are drawn orange. Finally, a person is drawn red if they were in the wrong position for operating table interaction.

Shown video frames were selected by the authors to demonstrate algorithm successes and failures. Keypoints with a detection confidence below 0.2 were not drawn. For each pose, an ID and a detection confidence score are shown.

6.2. RESULTS

6.2.1. DATASET



(a) Recording counts

(b) Recording time

(c) Recording durations per procedure

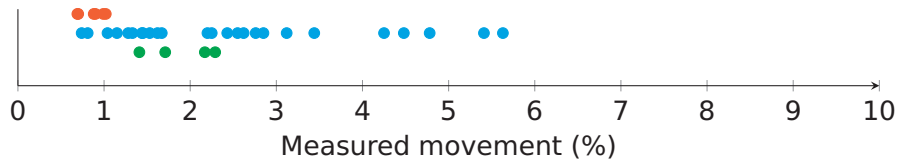
Figure 6.2.: Recorded procedure types and durations.

The dataset contains RAS, MIS, and OS procedures in the quantities shown in fig. 6.2a. Lighting-dependent framerates range from 6.2 to 26.1 frames per second (fps) during RAS and 12.5 fps to 25.8 fps during MIS. Since OS is performed with the lights on, the framerate was more constant here: from 24.7 fps to 25.3 fps. Recording durations are summarised in figs. 6.2b and 6.2c.

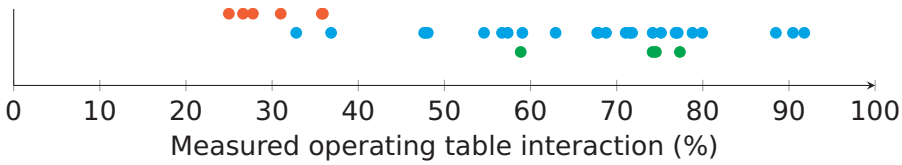
6.2.2. OPERATING TABLE INTERACTION OVER THE FULL DATASET

Measured over the entire dataset, the provided algorithm deems personnel to interact with the operating table 30.1 % of their time during RAS, 70.3 % during MIS and 70.4 % during OS. The algorithm classifies personnel as moving 0.8 % of the time during RAS, 2.0 % during MIS and 2.1 % during OS. Annotations report 24.6 %, 38.3 % and 37.6 % patient interaction during RAS, MIS and OS.

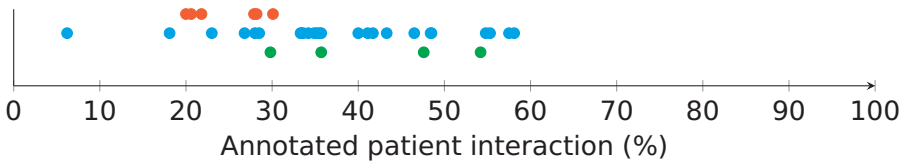
6.2.3. OPERATING TABLE INTERACTION PER VIDEO



(a) Measured movement



(b) Measured operating table interaction



(c) Annotated patient interaction

Figure 6.3.: Mean time fractions of movement, measured operating table interaction and annotated patient interaction per video per procedure type.

Figures 6.3a and 6.3b show measured movement and operating table interaction per video. The largest spread is seen between MIS

procedures, which range from 0.8 % to 5.6 % movement- and 32.8 % to 91.8 % operating table interaction time. RAS shows the least movement and operating table interaction time, from 0.7 % to 1.0 % and 25.0 % to 35.9 % respectively. OS has movement between 1.4 % and 2.3 % and operating table interaction from 58.9 % to 77.3 %. Annotated patient interactions per video in fig. 6.3c range from 20.0 % to 30.1 % during RAS, 6.2 % to 58.1 % during MIS and 29.8 % to 54.2 % during OS.

Two RAS procedures can be seen to have measured operating table interaction time fractions of 35.8 % and 35.9 %, whereas the rest scores only up to 31 %. During one of these, closing the entry wounds took an hour, whereas normally it takes about 15 minutes. As fewer people are near the table during RAS surgery than wound closure, more operating table interaction is detected during the latter.

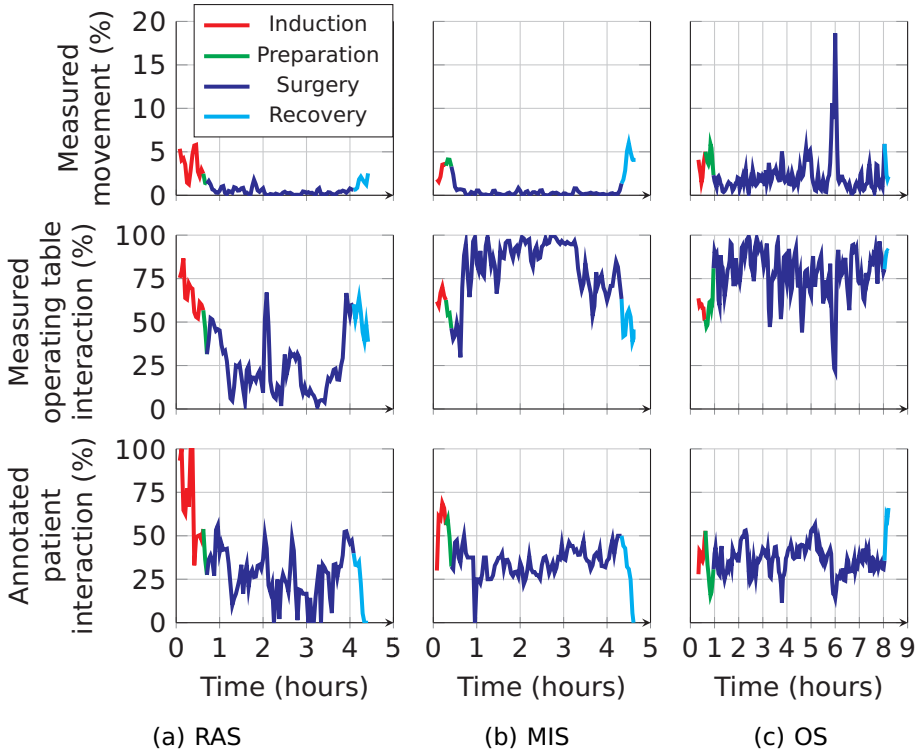
Looking at MIS, two procedures show operating table interactions of 32.8 % and 36.9 %, the others scoring at least 47.4 %. During one of these, two spectators are visible and detected during the entire procedure. The personnel at the operating table is poorly visible, due to the patient blanket having the same colour as their clothes. Three other procedures show 88.5 %, 90.4 % and 91.8 % measured operating table interaction. One of the three is a procedure with no spectators and with few people present beside those at the operating table. During another—again without spectators—a complication caused the surgery to take longer than the other workflow phases.

During OS, we observe the opposite as described in the previous paragraph. Here, most interaction with the operating table is observed during surgery. One procedure shows operating table interaction 58.9 % of the time, which is at least 74.2 % for the others. This procedure has a relatively long anaesthetic induction phase, spanning about one quarter of the recording. As opposed to surgery, few people are around during induction, and preparations are made in parallel throughout the room.

6.2.4. INTERACTION OVER TIME

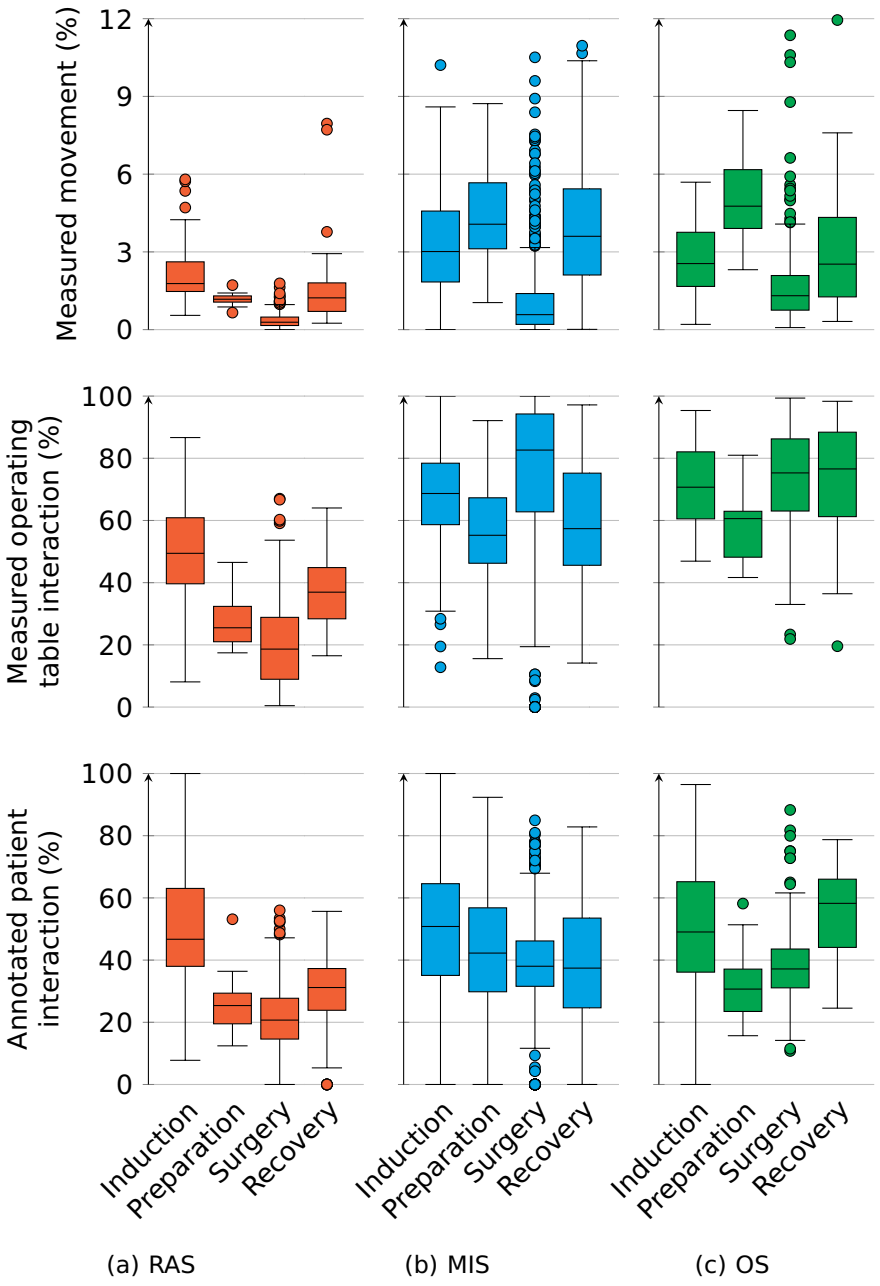
Figure 6.4 shows movement, measured operating table interaction, and annotated patient interaction over time during example RAS, MIS and OS procedures. During RAS and MIS, least movement is seen during the surgery phase. MIS and OS show the highest measured operating table interaction during this phase. Annotated patient interaction fluctuates around 40 % for all procedures. OS shows most measured variation throughout the procedure.

After six hours, the OS procedure shows a movement- and operating table interaction spike where the surgical team transitions from surgery to closure of the wound. The RAS procedure shows a spike in operating table interaction at two hours, where a robot arm was replaced. shortly thereafter another increase signals manual repositioning of a



(a) RAS (b) MIS (c) OS

Figure 6.4.: Measured movement and operating table interaction, and annotated patient interaction, over time during an example RAS, MIS, and OS procedure.



(a) RAS (b) MIS (c) OS

Figure 6.5.: Boxplots of movement, measured operating table interaction, and annotated patient interaction, during different phases and procedure types. Each datapoint is a time window like in fig. 6.4. Outliers are present beyond the y axis. Note that the range on the y axis differs between some subplots. Since datapoints are sampled in time using overlapping windows, measurements are not independent of each other.

robotic arm. The observed spikes are present in the annotated patient interaction during RAS but not during OS.

Measured movement over time per procedure type and phase is summarised in the top row of fig. 6.5. Less movement was measured in RAS procedures than other types. During the surgery phase of OS, the peak at six hours in fig. 6.4c shows up as an outlier. Less movement is measured during surgery than during the other phases for all procedure types. During RAS a lower median movement is measured during preparation than induction, whereas for MIS and OS this is the other way around. median movements during induction and recovery lie at most 59 percentage point apart for all procedure types.

The second and third rows of fig. 6.5 show measured operating table interaction and annotated patient interaction. Less table- and patient interaction are measured and annotated during RAS than other procedure types. The median annotated patient interaction lies higher during induction and recovery than preparation and surgery for RAS and OS procedures. During MIS most operating table interaction is measured during surgery, and least patient interaction is annotated during recovery. The main results from sections 6.2.2 to 6.2.4 are summarised again in table 6.6.

6.2.5. DETECTED AND ANNOTATED POSES

Figure 6.6 summarises the number of detected poses as a percentage of the annotated number of people over time. This was done separately for persons who were annotated and measured as interacting with the operating table, and those who were not. The median number of detected poses is always below 100 % for people not interacting with the operating table. For those who interact, detection percentages are higher in most cases. The difference between interacting and non-interacting detection percentages is larger for MIS and OS than RAS. The interquartile spread is also larger for interacting- than non-interacting persons.

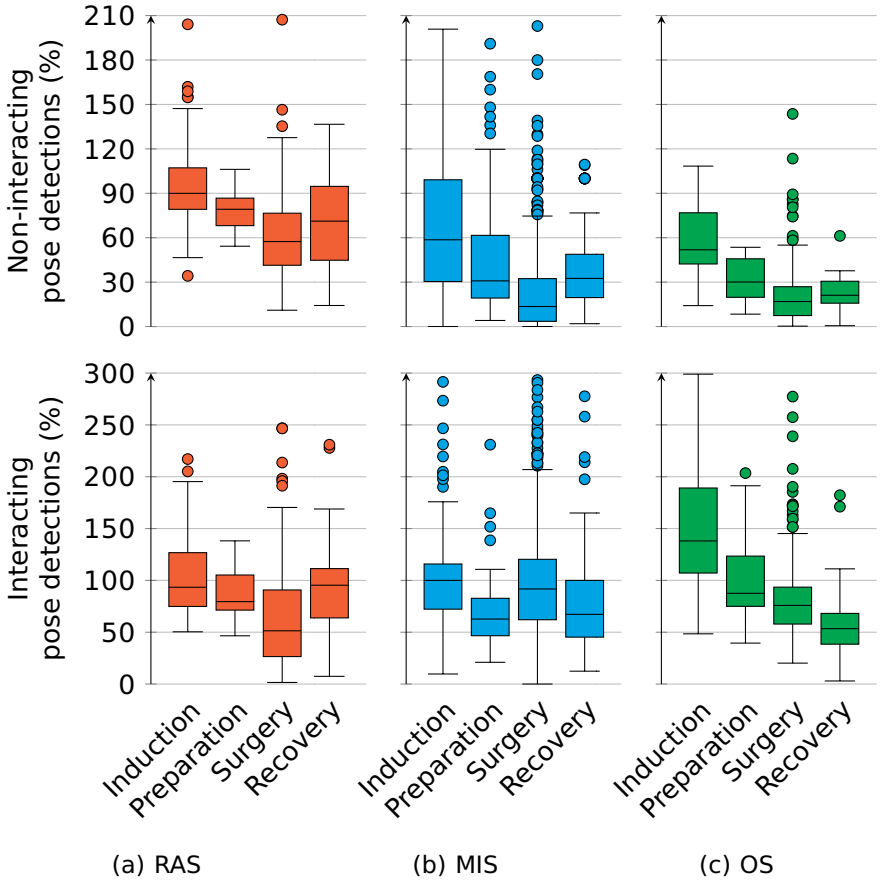
A larger fraction of non-interacting persons was detected during RAS than during MIS and OS. For interacting persons, detection percentages were more equal between procedure types. The least non-interacting persons were detected during the surgery phase for all procedure types. For interacting persons, this is the case only during RAS. As individual pose detections cannot be verified without annotating person locations, false negatives and false positives might nullify each other in the results of fig. 6.6.

6.2.6. QUALITATIVE RESULTS

A sample of human pose detections is shown in fig. 6.7. The top-left image shows three correctly detected poses. The cleaning person on the

Table 6.6.: Summary of the main results from sections 6.2.2 to 6.2.4. Means (%) are reported with one standard deviation (percentage point) where applicable. Note that, as RAS was recorded using a different camera system than MIS and OS, these results cannot be compared directly.

	Dataset	Per video	Per window			
			Induction	Preparation	Surgery	Recovery
Robot-assisted surgery						
Table interaction	30.1	30.3 ± 4.3	50.3 ± 15.0	28.0 ± 8.2	20.4 ± 13.8	37.2 ± 11.5
Movement	0.8	0.9 ± 0.1	2.2 ± 1.1	1.2 ± 0.2	0.4 ± 0.3	1.6 ± 1.5
Patient interaction	24.6	24.8 ± 4.1	49.6 ± 20.5	26.3 ± 9.6	21.4 ± 11.4	27.1 ± 14.8
Minimally invasive surgery						
Table interaction	70.3	66.2 ± 15.4	67.3 ± 17.5	56.0 ± 16.3	76.0 ± 22.9	59.6 ± 19.1
Movement	2.0	2.5 ± 1.4	3.4 ± 2.2	4.3 ± 1.8	1.5 ± 4.4	3.9 ± 2.5
Patient interaction	38.3	37.7 ± 12.3	47.6 ± 27.0	42.2 ± 20.4	38.9 ± 14.6	37.9 ± 21.9
Open surgery						
Table interaction	70.4	71.2 ± 7.2	71.9 ± 13.9	58.5 ± 10.9	73.7 ± 15.5	72.4 ± 20.6
Movement	2.1	1.9 ± 0.4	2.7 ± 1.4	5.0 ± 1.5	1.7 ± 1.8	3.8 ± 4.6
Patient interaction	37.6	41.8 ± 9.6	50.3 ± 21.9	31.9 ± 11.6	38.6 ± 11.2	55.3 ± 14.9



(a) RAS (b) MIS (c) OS

Figure 6.6.: Number of detected poses divided by the number of annotated persons in the room. The calculation was done separately for persons interacting- and persons not interacting with the operating table. Each datapoint is a window like in fig. 6.4. Outliers are present beyond the y axes.

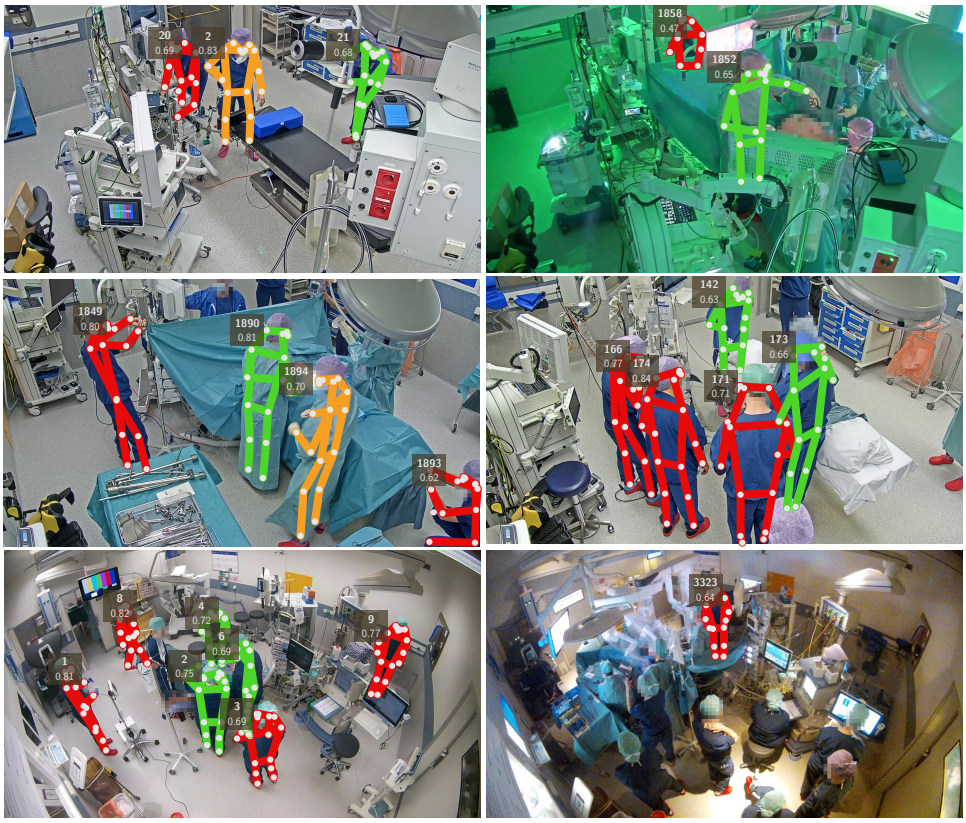


Figure 6.7.: Qualitative pose detections. Poses are drawn in green when classified as interacting with the operating table, orange when they are in the right position but moving too fast to interact, and red when they are in the wrong position. A tracking ID and detection confidence are shown for each pose.

right satisfies the positional and movement requirements, although their activity would not be classified as interacting with the operating table by a human. The person in the middle is not classified as interacting because of their speed. On the left, someone is busy elsewhere, whose left ankle was detected in the wrong place.

The top-right image shows ongoing MIS with the lights turned off. Only two out of five persons are detected. Hips and knees of the person in front are placed despite being occluded. Three staff members near the table are heavily occluded or face away from the camera, and are not detected. The shown IDs of 1852 and 1858 mean that the algorithm assigned and lost IDs 1856 times before this frame.

The middle-left image shows an OS procedure. Persons in front of the camera are detected with confidence scores of at least 0.70, with the exception of the partially out-of-frame person in the lower right corner. The surgeon is classified as interacting with the operating table, one assistant is in the wrong position for this, and another is turning away to move towards the instrument table. Three people in the back are not detected, each of which is either occluded, partially out-of-frame, or both.

The middle-right image was filmed during the induction phase of an OS procedure. Two out of the five detected people are close enough to the operating table to be classified as interacting. Three persons were not detected, all of which are occluded by clothing or another person, or partially out-of-frame.

The larger-field-of-view camera filming the RAS procedures makes persons appear smaller, as can be seen in the fifth image. All persons but one—who is occluded by the IV—are detected with a confidence of at least 0.69. The last image shows a later stage of the same procedure, with the lights off. Only one of the twelve persons is detected here. Looking closely, the sensor noise increased with respect to that when the lights were on.

6.3. DISCUSSION

In this work we quantified the interaction of personnel with the operating table, by analysing monitoring footage from 35 gynaecological procedures of three technology levels. Personnel movement was measured and interaction with the patient annotated, for a multimodal comparison between workflow phases and procedures of varying technology levels.

Annotated patient interaction suggests less interaction with the patient during RAS than other procedures. This could be caused by the nature of the procedure: personnel is spread through the room during RAS whilst the robot is interacting with the patient, and focussed around the operating table and patient during MIS and OS. Measured operating table

interaction shows a similar trend, although this result is biased by the differing camera systems and higher-quality pose detections near the operating table. Operating table- and patient interaction as a function of time differ similarly between procedure types as well. These differences are more pronounced in measured operating table interaction than the annotated patient interaction. Again, biases from differing camera systems and pose detection quality will amplify measured differences between RAS and the other procedures.

Section 6.2.5 suggests that a similar percentage of poses is detected during all procedure types. This view could be distorted, as there are more spectators—who are not annotated—during RAS than during MIS and OS. Qualitative results reflect this: since many false negatives here are unannotated spectators, the relative number of pose detections remains high. Detecting spectators reduces measured interaction without affecting the used pose detection metric. Similar reasoning applies to false positives, when persons are detected where there are none. Spectators and false positives explain the detection rates above 100 % in section 6.2.5.

Lights being off during RAS and MIS causes varying pose detection rates within these procedures.The built-in compensation from section 6.1.5 might not be sufficient with false negatives. When comparing results between workflow phases, this needs to be kept in mind.

Least movement is detected during the surgery phase. Here, most persons are busy at the table and spectators are standing still. The other phases show more movement variation, as preparations or cleanup are ongoing throughout the room. Most interaction is detected and annotated during induction or surgery—depending on the procedure type.

The algorithm and annotations measured different kinds of interaction by considering different properties of motion. Patient interaction was annotated based on observed intent and actions, and operating table interaction using only position and displacement of detected poses. Future algorithms could try to capture patient interaction using human action recognition. Here, nuances in intent should be taken into account. For instance, is waiting by the operating table to carry out a task an interaction, or is it idling? Are controlling the robot and monitoring the patient vitals technical or clinical tasks? When looking per procedure or procedure type, patient interaction was lower during RAS and MIS than OS. Within individual procedures, interactions with the operating table and patient evolved differently. For example, during OS, there was interaction with the patient, but not with the operating table, when personnel transitioned from surgery to wound closure.

Large fluctuations are visible in measured operating table interaction, where certain events or workflow phase transitions occur. These events are also visible in the annotated patient interaction, albeit to lesser

extent. Hence, the proposed measuring approach might prove valuable for workflow recognition purposes.

This work presents a first step in quantifying time spent on different activities in the OR. In future work, 3D pose detection could be used in the algorithm, which is less dependent on the used camera system [21]. This would mitigate perspective and occlusion issues. A pose detection algorithm should be used that is robust to motion blur and sensor noise in low-light conditions [22]. It should be refined for the OR by e.g. domain generalisation [23] using perioperative monitoring footage like MVOR [14]. A tracking method should be used that corrects for variable framerates. A scalable method, in addition to tracking poses robustly, should not rely on new operating table annotations in each OR. Instead, an object detector could be designed to locate the table automatically. The use of 3D poses solves perspective dependencies, removing the need to annotate or detect separate regions per subpose. The patient interaction annotations in this work could be replaced with a separate classification algorithm. Classifying patient interaction will likely require refined personnel features beyond position and movement, such as roles or action recognition [24].

Recognising the nature of personnel actions can play a role in context-aware systems for improved workflow or staff deployment. The algorithm indicated workflow events and anomalies, which can be used to streamline daily planning and care. For example, the turnover team could be notified when a procedure is finishing. Dashboarding workflow metrics could provide hospitals insight into their operation. This could help reduce expenses and improve workflow through well-informed decision making.

6.4. CONCLUSIONS

The presented algorithm is suitable to estimate high-level interaction with the operating table when used with a modern camera system. For lower-level analyses, a more descriptive input feature is necessary that is robust in OR conditions.

REFERENCES

[1] R. M. Butler, A. M. Schouten, A. C. van der Eijk, M. van der Elst, B. H. W. Hendriks, and J. J. van den Dobbelsteen. “Towards Automatic Quantification of Operating Table Interaction in Operating Rooms”. In: *Int. J. Comput. Assist. Radiol. Surg.* (2025). doi: 10.1007/s11548-025-03363-8.

[2] W. Zhang, H. Li, L. Cui, H. Li, X. Zhang, S. Fang, and Q. Zhang. “Research Progress and Development Trend of Surgical Robot and Surgical Instrument Arm”. In: *Int. J. Med. Robot. Comput. Assist. Surg.* 17.5 (Oct. 2021), e2309. doi: 10.1002/rcs.2309.

[3] A. M. Schouten, S. M. Flipse, K. E. van Nieuwenhuizen, F. W. Jansen, A. C. van der Eijk, and J. J. van den Dobbelsteen. “Operating Room Performance Optimization Metrics: a Systematic Review”. In: *J. Med. Syst.* 47.1 (Dec. 2023), p. 19. doi: 10.1007/s10916-023-01912-9.

[4] B. M. Gillespie, J. Gillespie, R. J. Boorman, K. Granqvist, J. Stranne, and A. Erichsen-Andersson. “The Impact of Robotic-Assisted Surgery on Team Performance: A Systematic Mixed Studies Review”. In: *Hum. Factors: J. Hum. Factors Ergon. Soc.* 63.8 (Dec. 2021), pp. 1352–1379. doi: 10.1177/0018720820928624.

[5] B. Zheng, E. Fung, B. Fu, N. M. Panton, and L. L. Swanström. “Surgical Team Composition Differs between Laparascopic and Open Procedures”. In: *Surg. Endosc.* 29.8 (Aug. 2015), pp. 2260–2265. doi: 10.1007/s00464-014-3938-3.

[6] J. Zamudio, J. Woodward, F. F. Kanji, J. T. Anger, K. Catchpole, and T. N. Cohen. “Demands of Surgical Teams in Robotic-Assisted Surgery: An Assessment of Intraoperative Workload within Different Surgical Specialties”. In: *Am. J. Surg.* 226.3 (Sept. 2023), pp. 365–370. doi: 10.1016/j.amjsurg.2023.06.010.

[7] S. S. Celik, Z. O. Koken, A. E. Canda, and T. Esen. “Experiences of Perioperative Nurses with Robotic-Assisted Surgery: A Systematic Review of Qualitative Studies”. In: *J. Robot. Surg.* 17.3 (June 2023), pp. 785–795. doi: 10.1007/s11701-022-01511-9.

[8] S. E. Lee, M. MacPhee, and V. S. Dahinten. “Factors Related to Perioperative Nurses’ Job Satisfaction and Intention to Leave”. In: *Jpn. J. Nurs. Sci.* 17.1 (Jan. 2020), e12263. doi: 10.1111/jjns.12263.

[9] M. F. H. Gil, J. A. R. Hernández, F. J. Ibáñez-López, A. M. S. Llor, M. D. R. Valcárcel, M. Mikla, and M. J. L. Montesinos. “Relationship between Job Satisfaction and Workload of Nurses in Adult Inpatient Units”. In: *Int. J. Environ. Res. Public Health* 19.18 (Sept. 2022), p. 11701. doi: 10.3390/ijerph191811701.

[10] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, M. Hashizume, D. Katic, H. Kenngott, M. Kranzfelder, A. Malpani, K. März, T. Neumuth, N. Padoy, C. Pugh, N. Schoch, D. Stoyanov, R. Taylor, M. Wagner, G. D. Hager, and P. Jannin. “Surgical Data Science for Next-generation Interventions”. In: *Nat. Biomed. Eng.* 1 (Sept. 2017), pp. 691–696. doi: 10.1038/s41551-017-0132-7.

[11] L. R. Kennedy-Metz, P. Mascagni, A. Torralba, R. D. Dias, P. Perona, J. A. Shah, N. Padoy, and M. A. Zenati. “Computer Vision in the Operating Room: Opportunities and Caveats”. In: *IEEE Trans. Med. Robot. Bionics* 3.1 (Feb. 2021), pp. 2–10. doi: 10.1109/TMRB.2020.3040002.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. “Microsoft COCO: Common Objects in Context”. In: *Eur. Conf. Comput. Vis.* Springer, Sept. 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[13] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. “2D Human Pose Estimation: New Benchmark and State of the Art Analysis”. In: *IEEE Conf. Comput. Vis. Pattern. Recognit.* IEEE, June 2014, pp. 3686–3693. doi: 10.1109/CVPR.2014.471.

[14] V. Srivastav, T. Issenhuth, K. Abdolrahim, M. de Mathelin, A. Gangi, and N. Padoy. “MVOR: A Multi-View RGB-D Operating Room Dataset for 2D and 3D Human Pose Estimation”. In: *Conf. Med. Image Comput. Comput. Assist. Interv.* MICCAI, 2018.

[15] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu. “AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 45.6 (June 2023), pp. 7157–7173. doi: 10.1109/TPAMI.2022.3222784.

[16] J. Redmon and A. Farhadi. *YOLOv3: An Incremental Improvement*. Apr. 2018. doi: 10.48550/arXiv.1804.02767. arXiv: 1804.02767v1 [cs.CV].

[17] K. He, X. Zhang, S. Ren, and J. Sun. “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 37.9 (Sept. 2015), pp. 1904–1916. doi: 10.1109/TPAMI.2015.2389824.

[18] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *Proc. 29th IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[19] R. M. Butler, T. S. Vijfvinkel, E. Frassini, S. van Riel, C. Bachvarov, J. Constandse, M. van der Elst, J. J. van den Dobbelsteen, and B. H. W. Hendriks. “2D Human Pose Tracking in the Cardiac Catheterisation Laboratory with BYTE”. In: *Med. Eng. Phys.* 135 (Jan. 2025), p. 104270. doi: 10.1016/j.medengphy.2024.104270.

[20] Y. Zhang, P. Sun, y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang. “ByteTrack: Multi-object Tracking by Associating Every Detection Box”. In: *Eur. Conf. Comput. Vis.* Springer, Oct. 2022, pp. 1–21. doi: 10.1007/978-3-031-20047-2_1.

[21] B. G. A. Gerats, J. M. Wolterink, and I. A. M. J. Broeders. “3D Human Pose Estimation in Multi-view Operating Room Videos using Differentiable Camera Projections”. In: *Comput. Methods Biomech. Biomed. Eng.: Imaging Vis.* 11.4 (2023), pp. 1197–1205. doi: 10.1080/21681163.2022.2155580.

[22] S. Lee, J. Rim, B. Jeong, G. Kim, B. Woo, H. Lee, S. Cho, and S. Kwak. “Human Pose Estimation in Extremely Low-Light Conditions”. In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2023, pp. 704–714. doi: 10.1109/CVPR52729.2023.00075.

[23] Z. Wang, R. Butler, J. J. van den Dobbelsteen, B. H. W. Hendriks, M. van der Elst, and J. Dauwels. “Towards Robust Object Detection in Unseen Catheterization Laboratories”. In: *IEEE Int. Workshop Med. Meas. Appl.* IEEE, June 2024. doi: 10.1109/MeMeA60663.2024.10596906.

[24] H. Kaur, V. Rani, and M. Kumar. *Human Activity Recognition: A Comprehensive Review*. Nov. 2024. doi: 10.1111/exsy.13680.

DISCUSSION

This chapter concludes the book. After discussing the answers to each subquestion separately, it treats the main research question. Several recommendations are made concerning the technical design of workflow monitoring systems. For a succesful implementation, it is important to consider applications, ethics, and stakeholders. The book is concluded with an epilogue after this chapter, which shares final thoughts of the author on future directions in workflow monitoring.

This thesis investigated human poses detection in interventional environments, and their use for the extraction of context-aware workflow information. In chapter 1, a research question and five subquestions were posed. Each chapter answered one subquestion. Section 7.1 summarises each of these answers separately. The rest of the chapter addresses the research question.

7.1. SUBQUESTIONS

This section summarises the answer to each subquestion from chapters 2 to 6.

7.1.1. WHICH COMPUTER VISION TECHNOLOGIES CAN BE APPLIED FOR CATH LAB MONITORING?

Chapter 2 investigated a computer vision pipeline in the cardiac catheterisation laboratory (Cath Lab) and operating room (OR). Promising technologies were object detection and pose detection. Object detectors find object bounding boxes in images. Pose detectors find object- or human keypoints that describe their pose. Due to varying object appearances, object detection did not generalise to different Cath Labs. Pose detection, however, generalised well because human anatomy remains similar in different environments. Here, improvements could be made by training algorithms on persons wearing sterile Cath Lab- and OR garments.

Computer vision methods struggled with e.g. occlusion and reflections in the Cath Lab. These effects were mitigated using multi-view information. Cameras were calibrated to relate information between viewpoints. Cameras occasionally move between procedures, and placing permanent calibration patterns was no option. Therefore, calibration was performed using known keypoints of stationary objects. Confirming object detections in one view using information from others yielded no significant improvement. For pose detection, camera calibrations were succesfully used to triangulate 3D poses that were robust agains occlusion.

7.1.2. WHICH 2D HUMAN POSE ESTIMATOR PERFORMS BEST WITHIN THE VISUAL COMPLEXITY OF THE CATH LAB?

Chapter 3 decided to continue with 2D poses. Objects detection did not generalise well, and 3D poses were limited by flawed camera synchronisation and calibration. Several state-of-the-art human pose detectors- and trackers were benchmarked in the Cath Lab: AlphaPose, OpenPifPaf and OpenPose. This yielded a tradeoff between metrics,

where OpenPifPaf gave the best keypoint locations, and AlphaPose more accurate full poses and confidence scores.

Results varied by viewpoint, where pose detectors struggled considerably with viewpoints that see occlusion and reflections. In addition, humans were often not detected when facing away from the camera. Due to sterile clothing, some human keypoints—the shoulders, hips and knees—were often placed inaccurately and with low confidence.

Person reidentification (ReID) did not perform well. The tracking version of AlphaPose uses human appearance for ReID. However, because everyone in the Cath Lab was dressed the same, this resulted in identity swaps. OpenPifPaf did not do better. The underlying tracking technology caused this method to merge individuals, by connecting the wrong keypoints.

7.1.3. How to reidentify persons in the Cath Lab?

As human pose trackers in chapter 3 did not perform well in the Cath Lab, chapter 4 investigated an alternative method. Since a major issue was the similar appearance of different persons, no visual information was used for ReID. Instead, tracking was done purely using geometric information. The BYTE algorithm was adapted to use poses instead of bounding boxes, and take into account several orders of motion. Besides ReID, this algorithm refined keypoint positions based on tracking information.

The resulting algorithm was faster and more accurate than those tested in chapter 3. The inclusion of pose information over bounding boxes, and motion beyond the first order, barely contributed. After prolonged occlusion, the tracker could not ReID persons because visual information was not considered. Therefore, the suggested algorithm was suitable to track short-term motion only.

7.1.4. Which aspects of motion are most descriptive of Cath Lab workflow?

Chapter 5 investigated the presense of workflow information in different aspects of human motion. Learning about the contribution of different motions and interactions may provide the insights necessary for constructive feedback applications. An algorithm was designed to classify workflow phases based on handcrafted features—representing relative personnel position and motion. Classification was done separately on each feature for interpretability. The resulting algorithm was not accurate, or suitable for practical applications. However, the flawed results did suggest recommendations for future algorithm designs.

The distance between ankles of different people was an important feature, as it seemed to indicate procedure progression. When

including actual procedure duration as a feature, distances between individuals became less important in general. Features concerning keypoint relations within a person benefited when considering duration. Especially the upper body keypoints seemed informative in this case. Relations between counterintuitive keypoint pairs turned out to be descriptive, such as the location of the shoulders with respect to the nose.

The proposed algorithm lacked association between different features. In future algorithms, more association can lead to better reasoning within the classifier, and hence better predictions. Designing a classifier, relations between all keypoints within an individual should be considered for feature extraction. Spatial relations between different individuals may be less important, as long as procedure duration—or another measure of progression—is included as a feature. Besides added association, prior knowledge could get a more prominent role in such algorithms. If the role of each person can be detected, this could enhance workflow information. A distinction could be made between motion of the patient, cardiologist, and assistants for workflow analysis. A carefully designed algorithm may provide context about the reasoning behind a classification, which can be used in feedback applications.

7.1.5. How can personnel actions be classified in the OR?

Rather than considering complex motion, chapter 6 aimed to recognise staff interaction with the OR operating table. Measuring differences in interaction between different kinds of surgery yields insights into the effects of new technologies on interventional workflow. This is especially important because medical personnel reports higher perceived workloads and stress, resulting from technologies such as robotic-assisted surgery. Classification was based purely on personnel position and displacement. 2D pose data enabled correction for the camera perspective. The suggested method simply checked whether persons stand still near the table.

Less interaction was measured during robot-assisted surgery than minimally invasive- and open surgeries. The results were skewed by the presense of spectators and flaws in the camera system. Differences were detected between workflow phases. During surgery, people moved the least but interacted with the table the most. High interaction was measured during anaesthetic induction as well, and most movement during preparations and cleanup. Although interaction measurements and annotations showed differences, this was partially explained by differences in definition. Both measurements and annotations showed similar interaction trends between procedures or phases. Some workflow events produced patterns in the algorithm output, but were not visible in annotations.

The main research question is discussed in the remainder of this chapter:

"How can we extract workflow information from Cath Lab monitoring video footage, and use it to improve interventional efficiency?"

7.2. FEATURE EXTRACTION

The first step towards workflow analysis from monitoring videos is to extract descriptive information. In this thesis, several features were considered in terms of usefulness to workflow analysis, and practicality of implementation.

Object- and human pose tracking provide the whereabouts of important actors over time. The motion of individual actors is descriptive to their current action. Combining individual actions, and interactions between multiple actors, provides information about group activities such as workflow phases.

The Cath Lab and OR are complex environments for computer vision. For example, sterile clothing, hairnets and masks make staff hard to detect. Everyone dressing the same complicates ReID, and loose clothing and shielding occlude bodyparts. When facing away from a camera, individuals are often not detected. Tracking difficulties can be mitigated by providing different sets of clothing, e.g., in different colours. Additionally, markers could be printed on the clothes to enable marker-based pose detection. Since privacy is a big issue in healthcare, no identification method can interfere with privacy regulations.

In Cath Labs, the many objects and persons in a small space often occlude each other or themselves. When cameras are mounted on the ceiling, occlusion is especially present for objects and bodyparts near the floor. Ceiling-mounted arms that carry e.g. the mobile lead shield, can be positioned by accident to occlude entire viewpoints. Occlusion is even more present in the OR, with more persons and equipment than in the Cath Lab. When objects and persons are not detected due to occlusion, algorithms are blind to the workflow information they represent.

Specular surfaces such as displays and windows present another challenge. Current computer vision algorithms cannot distinguish between real objects and reflections. Hence, an algorithm may detect an object where there is none, receiving false workflow information. Reflections are not as visible when the lights are off, which is the case during minimally invasive surgeries.

Many algorithms that rely on temporal features assume a constant framerate. Some phases of minimally invasive procedures are carried out in the dark, to which cameras adapt their exposure times. This introduces motion blur and sensor noise, and object- and person

appearances change with lighting conditions, complicating detection. Detection algorithms for the Cath Lab and OR should be trained in such varying lighting. Varying frame times can disturb the motion prediction used in many tracking algorithms. Hence, algorithms relying on temporal features should correct for the none-constant frame times. On the upside, reflections are not as present in the dark as with the light on.

Both occlusion and reflections depend on the camera viewpoint. Objects that are occluded or reflected from one perspective, may not be from another. Therefore, it is beneficial to incorporate information from multiple viewpoints in feature extraction. Combining multi-view features or implementing voting systems can improve feature quality through redundancy.

To relate information between viewpoints, periodic calibration is needed in case the cameras are moved. It is not practical to place permanent calibration patterns in Cath Labs and ORs. Hence, a method should be used that performs pattern-less extrinsic calibration, e.g. using keypoint detections.

A downside of multi-view camera systems is that objects look different from different perspectives. Worse, object appearances vary in different Cath Labs and ORs. If a method is to be scaled, it is not practical to require new object annotations for every camera, room, and hospital. Therefore, a feature extraction method should be chosen that generalises well. This will be easier for pose- than bounding box detection. After all, human anatomy is similar everywhere, and large datasets are publicly available with annotated human poses.

Even if detected correctly, object position and motion are perceived differently from different viewpoints. For example, when an object moves to the right with respect to one viewpoint, an opposing viewpoint would observe movement to the left. Hence, conveyed workflow information will not be the same. In the worst case, workflow analysis has to be implemented separately for each new camera or room. A solution is to triangulate 2D features from multiple viewpoints into 3D features. 3D features are invariant to camera positions, as long as the camera system is installed appropriately as discussed in section 7.3.

Features besides monitoring video can enhance robust context-awareness. For instance, some systems log their usage, sensors measure door movements, laparoscopic video is recorded during minimally invasive surgery, and X-ray video during Cath Lab procedures. Such features are room-invariant, e.g., a door sensor output looks the same regardless the door appearance and position. Incorporating such robust features creates a multimodal approach, where data that are not (easily) extracted from video monitoring are made available. Additionally, richer video data can be recorded. For instance, cameras can be mounted that measure distance in addition to colour.

7.3. MONITORING SYSTEM DESIGN

The datasets used in this thesis were recorded using various camera systems. Besides the visual complexity of the Cath Lab and OR, camera system limitations presented a major challenge to feature extraction. This section reviews the encountered camera system flaws, and makes recommendations on the design of future systems.

The most important feature of a camera system is how much of the room is covered. At the very least, the area surrounding the operating table should be in view. For full context-awareness, the entire room should be visible in the combination of all viewpoints. Room coverage can be enlarged in two ways. The first is to use cameras with a large field of view, e.g. a 'fish eye' lens. The larger the field of view, however, the more distortion is introduced in recordings. Distortion complicates feature extraction, and especially camera calibration. The second option is to add more cameras with different perspectives. The more cameras are hung, the less sensitive the overall system will be to reflections and occlusion. With sufficient overlap between their perspectives, multiple cameras enable multi-view- and 3D feature extraction. However, the more cameras are present, the more computationally expensive it will be to process their video output. Thus, real-time applications will impose an upper bound on the number of cameras, recording resolution, and/or framerate. A tradeoff between field of view and the number of cameras should be based on the room and application.

More nuance is present in camera placement besides full-room coverage and viewpoint overlap. Cameras can be mounted on- or near the ceiling to get the best overview of a room. Here, however, they also have the largest risk of being occluded by mounting arms. In addition, their distance from the procedure means that detected positions will be less precise. If one aims to record, e.g., individual finger movements, it is better to mount cameras lower and/or closer to the procedure. Here, the recordings will show more detail, but will suffer more occlusion from e.g. personnel walking by.

Besides camera placement and lenses, sensor specifications play a role. High resolutions and framerates contribute to recording detail, although they do take a toll on computational processing requirements. A camera sensor should be chosen by performance in varying lighting conditions, adapting quickly to changes and limiting noise and framerate drops in low light. If it does not interfere with equipment in the room, cameras could carry infrared lights to keep the perceived lighting more constant.

A vital aspect of multi-view feature extraction is synchronisation. When multiple cameras record a frame simultaneously, they need to reliably record the same moment. If different cameras record with a delay of seconds or minutes from each other, their information cannot be related properly. Multi-view- or 3D features cannot be extracted in this

case. If framerates are adapted to lighting, all cameras should do this in the same way at the same time. Some cameras support synchronisation by e.g. receiving recording signals through a wire. If synchronisation is lacking, recordings should be synchronised in postprocessing, based on recorded metadata and/or visual features.

7.4. ANONYMISATION AND PRIVACY

A downside of monitoring is the privacy-sensitivity of video data, which show patients and staff. Such videos should be handled with utmost care, as a leak will have serious repercussions for the recorded individuals and others involved. A procedure should never be recorded without the explicit, contractual informed consent of all individuals in the room. The rest of this section describes considerations in video data handling to protect the privacy of filmed subjects.

In the datasets presented in this thesis, faces were blurred to protect the identities of those present. Done manually, the necessary face annotating is a labour-intensive process. Faces can be detected automatically using computer vision instead. However, this only works if the face detection is reliable. In practice, computer vision algorithms were not yet reliable enough to accurately detect all faces in perioperative monitoring video. Even if faces are blurred perfectly, persons could be identied by e.g. bodytype, posture or way of moving. Moreover, blurring parts of the image erases workflow information and complicates bounding box- and pose detection. Therefore, blurring monitoring videos should be avoided if alternatives exist.

An alternative to blurring could be not storing the videos at all. During recording, features and/or workflow information can be extracted automatically, and stored instead of frames. This processing could be done within each camera for single-view features, or in a single computer for multi-view or 3D features. Possible anonymous features are the mentioned bounding boxes and poses. More informative, however, would be to store lower-level features such as the (intermediate) feature maps from a pretrained backbone neural network (NN). These maps contain visual information that can further be processed by computers, but is not easily interpretable by humans. By processing recordings immediately, no human needs see any raw videos. A limitation of storing anonymous features only, is that it may complicate the application of feedback systems from section 7.5. After all, if no information remains about identity, it will be impossible to give personalised feedback to those who could benefit.

Besides recording, the storage and transportation of data are a sensitive topic. Privacy-sensitive data are not allowed to leave their hospital of origin, which complicates open research. This slows down developments in medical (workflow) technology: to enable some

products or research, publishing data or centralised storage may be necessary. After preprocessing as described before, the storage and transportation of completely anonymous features may be less of an issue. Data may even be published in the open domain. For such plans, it needs to be verified that there is absolutely no way of retrieving personal information from the shared features. In other words, the used feature extractor must have no possible inverse, or be kept secret.

Workflow analysis leads to the extraction of workflow metrics, possibly on an individual basis. These metrics could be interpreted as a measure of personal performance. It is important to consider who gets access to these metrics, and the identity of those they belong to. If each individual receives their own metrics only, this could yield a fruitful feedback system. If metrics per individual are accessible by employers, other employees, or the public, this could potentially yield undesirable situations. Systems that are designed to help people, could be misused or misinterpreted to impose judgement. This dystopic working environment of fear is a reason for medical professionals today to distrust monitoring systems and workflow research. To prevent such situations, serious thought should go into the question who gets access to which information. A workflow optimisation system should work constructively, without posing a threat to any stakeholder.

By itself, workflow sensing serves no purpose. Such sensing only becomes useful when the acquired knowledge is applied to improve situations for the patient, staff or hospital. Some possible applications are described in the next section.

7.5. APPLICATIONS

Workflow detection and context-awareness can fuel applications in several directions. The work presented in this dissertation provides a basis for the scalable extraction of workflow metrics from monitoring videos. These metrics could be presented to the individual concerned, to provide insights into their way of working. This could be done in the form of a dashboard. Such a dashboard could show e.g. the time spent per phase, performance per patient type, X-ray exposure, and comparisons to historical metrics and (anonymous) peers. If analysis is centralised, different ways of working can be accumulated and compared in terms of outcome. For instance, different sequences of (parallel) workflow steps can be linked to procedure success rates, phase durations, and employee wellbeing. After procedures, staff could receive feedback or suggestions to improve their workflow, based on measurements from hospitals worldwide. It could be suggested to change the order in which tasks are performed for efficiency, or the way in which they are performed for better ergonomics. In addition, specific precautions could be recommended against X-ray exposure. Newfound information

on optimal workflow could be used directly in the education of new personnel.

Besides feedback and education, robust context-awareness could enable automated assistance or recommendations in real time. This way, technology collaborates directly with the staff. If a staff member is estimated to receive high radiation, the system could recommend them to stand elsewhere or move the lead shield. If a system sees a new workflow phase approaching, it could suggest new settings for the C-arm or X-ray. After procedures, the system could automatically fill out forms or summaries instead of the staff. In planning, medical teams could be suggested with employees who collaborate effectively, and linked to patients they will likely be able to help best. When a procedure is finishing, planners could be notified automatically to signal e.g. preparation of the next patient. In all these cases, the suggestions should be approved by the corresponding professional, before the system carries them out.

There is still a large gap between the work from this dissertation and the described applications. This book focussed on techniques for reliable computer vision in ORs, and explorative workflow phase recognition. True contextual understanding of the scene is still far off. However, basic workflow analysis using human poses was explored in chapter 5 and succesfully demonstrated in chapter 6. Continuations in this line of research can allow a system to understand the scene, enabling the outlined applications. An advantage of the monitoring approach from this work is that information about the entire perioperative workflow can be measured. Other datastreams, e.g., laparoscopic videos or medical images, give a more limited perspective.

Although promising, computer vision in its current state is flawed. Artificial intelligence (AI) reasoning is in its infancy, and difficult to interpret by humans. AI is rigid, and cannot extrapolate to unexpected situations. AI cannot reliably learn on the job by itself. For all these reasons, at the time of writing it would be unwise to let AI have true responsibility in healthcare. Instead, it should be seen purely as a set of useful tools. AI can provide advice and minor assistance, as long as it does not take action without the explicit approval of healthcare professionals. Full automation, e.g., an autonomous C-arm or even a robotic surgery platform, might become reality, but only in the far future.

7.6. BENEFICIARIES

The success of workflow optimisation should be measured in benefits to the stakeholders. Stakeholders include the patient, staff, hospitals, and societies as a whole. New technologies should support those involved, rather than getting in the way. For example, as noted in chapter 6,

robot-assisted surgery is currently often quoted by staff to yield more stress and delays. New technologies should have a positive effect on stakeholders, without introducing new problems. This section discusses some desired effects of workflow optimisation, for those who it directly affects.

In healthcare, patient wellbeing is the absolute priority. Besides their physical problem, patients are reported to deal with emotional discomforts such as anxiety for the procedure or annoyance over delays. This, in turn, affects healthcare professionals emotionally. The applications from section 7.5, enabled by context-aware systems from this and other studies, can streamline patient care. Education and automated assistance improve procedure efficiency and safety, reducing patient discomfort and waiting times.

The goal of perioperative personnel is to provide care. If changes are made to their working environment, these should be there to support them in that goal. The global shortage of 15.4 million healthcare professionals in 2020 causes stress and high workloads in the perioperative setting. New technologies should be low-profile, non-obtrusive, and intuitive to work with. The goal should be to improve the way of working and quality of care, whilst reducing the workload. This dissertation lays the groundwork for future context-aware systems, that provide staff with support and concrete handles to improve their workflow. Seeing a measured improvement in their performance, or receiving concrete recommendations, might be motivating and increase work satisfaction. Additionally, context-aware AI can assist with administrative tasks, allowing healthcare professionals to stay focussed on caregiving.

Hospitals aim to provide the best possible care within their resources. ORs especially represent a large part of hospital expenses. Data-driven optimised planning and efficient personnel allow for better care and reduced waiting times, within the same budget. In other words, this technology enables hospitals to deploy resources more efficiently or serve a greater population.

In 2020, 47 % of global healthcare was reserved to 22 % of the total population. Some societies have better access to healthcare than others. In developed countries, the population is aging; the number of patients outgrows the availability of personnel in the long term. To maintain health facilities, this shift must be compensated through e.g. support systems like the ones suggested in this thesis. In this work, the focus lies on human poses for scalable computer vision in ORs. Upon further development, the resulting scalable context-aware tools can easily be applied in any hospital.

EPILOGUE

Although process optimisation, medical technology, and computer vision are mature fields in academics and industry, the small area that combines the three remains in its infancy. In the rapid development of surgical data processing solutions, the cruciality of communication between engineers and clinicians is often overlooked. During my Ph.D. I chose to solve technical problems voiced by researchers and literature, or that I observed myself. This theoretical approach introduced a gap between my work and the practical needs of medical teams. In line with this, I learned in conversation that high maintenance and complexity lead to some medical technologies even hindering procedures. Conversely, I noticed that clinicians were often not aware of the potential and limitations of technology. A major theme here is the difference between human reasoning and computation: what is simple for a human can be very complicated for a computer, and vice versa. To engineers with experience in software or computer science this seems trivial, but for successful collaboration it must be remembered how it can be intrinsically counter-intuitive. This is especially true in fields where computers mimic human functioning, such as in computer vision. These examples illustrate the importance that engineers understand the needs of clinicians, and clearly communicate the possibilities of the current state-of-the-art.

Steve jobs made a remark in 1997:

"One of the things I've always found is that you've got to start with the customer experience, and work backwards for the technology."

I think that this remark fits perfectly in the healthcare technology sector today. Even with perfect computer vision in the OR, and cutting-edge context-aware reasoning systems built upon it, workflow analysis is useless without an application that the helps the stakeholders. Instead of extracting features, reasoning about workflow, and finding cool applications, this process should be reversed. The first step should be to talk to hospitals, staff and patients, find out what they need and what bothers them, conceive an application, and then reason what features need to be extracted.

Only then, should you monitor the contextual Cath Lab.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my promotors: John van den Dobbelsteen, Benno Hendriks and Maarten van der Elst. Our discussions were especially insightful thanks to the many backgrounds and perspectives on the table, and always yielded new ideas to advance the research. Whenever I got lost in details and theory, as I do, you never failed to guide me back to reality. Your honest and direct feedback ultimately proved vital for my personal and professional growth. You recognised my efforts to learn, explore and improve, and allowed me the freedom for it.

I am also incredibly grateful to my defence opponents: Robert van der Boon, Jenny Dankelman, Jouke Dijkstra, Danny Ruijters, and Thijs Vlugt. Thank you for spending time and effort to critically review this work.

The research in this dissertation was built directly upon the past efforts of several professionals. I want to express my appreciation towards Jan Constandse, Chavdar Bachvarov and Sjors van Riel, who arranged the camera system of chapters 2 to 5. Special thanks to Chris Varekamp for instructing me on the topic of camera geometry. Many thanks to Justin Dauwels, whose extensive knowledge enabled the adaptation of computer vision tools to the Cath Lab. I am very grateful to Anne van der Eijk, who provided crucial guidance and feedback for chapter 6.

I could not have finished this journey without my colleagues of the Medical Process Engineering group. I deeply appreciate the efforts of Teddy, who recorded the Cath Lab videos, explained medicine to me, and built pyramids with me in the Cath Lab for camera calibration. I am extremely grateful to Emanuele, with whom I spent hours annotating bodyparts, discussing mathematics, and in the spa when we should have been at a conference. My deep gratitude goes to Anneke; collaborating with you made the research and writing of chapter 6 not just possible, but truly a wonderful time and a treasured memory. Many thanks to Anton, who was always eager to discuss anything, drove halfway through Spain with me, and organised an amazing writing retreat.

Special thanks to my colleagues Pier, Mauricio, Katerina, Judith, Koen, Jonathan, Ioannis, Chun-Feng and Janne, you always made the fishtank feel like home (when I was there, anyway). I would also like to thank Hugo, Kirsten, Eline, Leila, Costanza, Alina, Andrea, Athina, Vitoria, José, Sietske, and Vishal, who formed the 3mE PhD council with me for a while. Many thanks to my colleagues Ebrahim, Marit, Suzanne, Niko, Jette, Monika, Merle, Robin, Bart, Nick, Martijn, Vera, Jan-Willem, Dirk, Jette, Esther, and all others of the BioMechanical Engineering group. You

made each conference a pleasure, and I seem to recall taking a boat trip with some of you (although I do not quite remember how that ended).

Many thanks to Peter and Teun, with whom I co-lectured a course and who guided me in the process. In addition, I have Barbara, Lennart, Boy, Charlotte, Myrthe and Sanne to thank for the opportunity to join the Hoe?Zo! science show. It was a pleasure to share this unique experience with Aafke, Aike, Ayla, Bodine, Janna, Lara, Maarten, Maartje, Merel, Pilar, Ralph, Roy, Sterre and Tim.

During my research I was involved in the supervision of several graduate students. I owe a great deal to Renjie, Yingfeng, Jinchen and Zipeng, whose works shaped chapter 2. Although I already forgot the little Mandarin you taught me, I will never forget the hours we worked side by side in the hospital. I would also like to mention Yuan, my first graduate student, who managed to graduate during COVID-19 despite the difficulties of this period. Many thanks to Hilda, Jasper, Tiara, Pepijn, Bardia, Thijs, Yannick and Carlijn, in whose projects I played an advisory role and/or whose insights helped my research forward. Finally, thanks to Femke and Quinten, whose annotations were vital for chapter 6. It was a true pleasure to work, learn, and chat with all of you.

Beside academic development, the promotional period was an important time of personal growth and self-reflection for me. I am indebted to Jan, Wendy, Isis, Ilse, Fedor, Cândida, Martin, Laura, Mariska and Alied for their indispensable guidance during this journey.

I have my friends to thank for making these years fun, exciting, and sometimes bearable. Special thanks to Thomas, who has been a lifelong friend and created the beautiful cover art for this book. Many thanks as well to Kadir, Chris and Danny, whose competition motivated me to finish my PhD first. I am incredibly grateful to Juda, who is a kind listener and helped me find myself. My gratitude also goes to Evelien, in whose home I am always welcome for tea, games, and talks. I am thankful to Esmée, who knows how to party, and how to keep up on our snowboarding trips. When I was too busy with the PhD to celebrate my birthday, the surprise party you organised truly touched my heart. My gratitude goes to Vinícius and Astrid, who showed me what a kick-ass PhD defence looks like. I would also like to thank Joris and Timothy, things are always fun with you around! I live a full life thanks to all of you, our adventures, sports, game nights, dinners, parties, trips, conversations, laughs and tears.

Lastly, I could not describe in words the love and support I received from my parents, Hans and Petra, and my brother Frank. I could not have finished this book without your continued support. Without your enthusiasm and motivation, I might never even have started it. Whenever I stumbled during the long road that led to this moment, you always pulled me back onto my feet. Finally, I am incredibly grateful for the support of my grandparents: Henk, Ank, Cor and Elma. You were always invested in my wellbeing and progress. You never lost faith.

CURRICULUM VITæ

Rick Maarten Butler

05-09-1997	Born in Eindhoven, the Netherlands.

EDUCATION

2009–2015	High School Christiaan Huygens College, Eindhoven, NL Atheneum, Natuur & Techniek
2015–2018	B.Sc. in Electrical Engineering *Cum Laude* Eindhoven University of Technology, NL
2018–2020	M.Sc. in Electrical Engineering *Cum Laude* Duke University, Durham, USA (2018–2019) Eindhoven University of Technology, NL (2019–2020)
2020–2025	Ph.D. in Mechanical Engineering Delft University of Technology, NL *Thesis:* Contextual Operating Room Monitoring: What pixels tell us about workflow *Promotors:* Prof.dr. J.J. van den Dobbelsteen Prof.dr. B.H.W. Hendriks Prof.dr. M. van der Elst

LIST OF PUBLICATIONS

FIRST AUTHOR MANUSCRIPTS & PRESENTATIONS

7. **R. M. Butler**, A. M. Schouten, A. C. van der Eijk, M. van der Elst, B. H. W. Hendriks, and J. J. van den Dobbelsteen. "Towards Automatic Quantification of Operating Table Interaction in Operating Rooms". In: *Int. J. Comput. Assist. Radiol. Surg.* (2025). doi: 10.1007/s11548-025-03363-8

6. **R. M. Butler**, E. Frassini, T. S. Vijfvinkel, S. van Riel, C. Bachvarov, J. Constandse, M. van der Elst, J. J. van den Dobbelsteen, and B. H. W. Hendriks. "Benchmarking 2D Human Pose Estimators and Trackers for Workflow Analysis in the Cardiac Catheterization Laboratory". In: *Med. Eng. Phys.* 136 (Feb. 2025), p. 104289. doi: 10.1016/j.medengphy.2025.104289

5. **R. M. Butler**, T. S. Vijfvinkel, E. Frassini, S. van Riel, C. Bachvarov, J. Constandse, M. van der Elst, J. J. van den Dobbelsteen, and B. H. W. Hendriks. "2D Human Pose Tracking in the Cardiac Catheterisation Laboratory with BYTE". in: *Med. Eng. Phys.* 135 (Jan. 2025), p. 104270. doi: 10.1016/j.medengphy.2024.104270

4. **R. M. Butler**, E. Frassini, T. S. Vijfvinkel, S. van Riel, C. Bachvarov, J. Constandse, M. van der Elst, B. H. W. Hendriks, and J. J. van den Dobbelsteen. "Heart Intervention Efficiency". Presented at: *Int. Soc. Med. Innov. Technol. Conf.* Sept. 2024

3. **R. M. Butler**, T. S. Vijfvinkel, E. Frassini, S. van Riel, C. Bachvarov, J. Constandse, M. van der Elst, J. J. van den Dobbelsteen, and B. H. W. Hendriks. "2D Human Pose Tracking in the Cardiac Catheterization Laboratory with BYTE". Presented at: *Comput. Assist. Radiol. Surg. Int. Congr.* June 2024

2. **R. M. Butler**, T. S. Vijfvinkel, S. van Riel, C. Bachvarov, J. Constandse, M. van der Elst, J. J. van den Dobbelsteen, and B. H. W. Hendriks. "2D Pose Tracking in the Cath Lab". Presented at: *Dutch Bio-med. Eng. Conf.* Jan. 2023

1. **R. M. Butler**, T. S. Vijfvinkel, J. Constandse, C. Varekamp, S. van Riel, C. Bachvarov, M. van der Elst, B. H. W. Hendriks, and J. J. van den Dobbelsteen. "Quantifying Cath Lab Workflow with 3D Poses". Presented at: *Int. Soc. Med. Innov. Technol. Conf.* May 2022

CO-AUTHOR MANUSCRIPTS

7. T. S. Vijfvinkel, E. Frassini, Y. Weijenberg, B. Nikookar, **R. M. Butler**, J. Constandse, M. van der Elst, B. H. W. Hendriks, and J. J. van den Dobbelsteen. "Automated Phase Recognition in the Cardiac Catheterization Laboratory with Deep Learning". Submitted for publication.

6. T. S. Vijfvinkel, **R. M. Butler**, T. P. Ringers, J. Constandse, V. W. Verhoeven, B. H. W. Hendriks, J. J. van den Dobbelsteen, and M. van der Elst. "Procedure-specific Exposure to Scatter Radiation during Cardiac Catheterizations: The Effects of Distance, Time and Shielding". Submitted for publication.

5. A. M. Schouten, **R. M. Butler**, C. E. Vrins, S. M. Flipse, F. W. Jansen, A. C. van der Eijk, and J. J. van den Dobbelsteen. "Impact of Operating Room Technology on Intra-operative Nurses' Workload and Job Satisfaction: An Observational Study". In: *Int. J. Nurs. Stud. Adv.* 8 (June 2025), p. 100341. doi: 10.1016/j.ijnsa.2025.100341

4. E. Frassini, T. S. Vijfvinkel, **R. M. Butler**, M. van der Elst, B. H. W. Hendriks, and J. J. van den Dobbelsteen. "Deep Learning Methods for Clinical Workflow Phase-Based Prediction of Procedure Duration: A Benchmark Study". In: *Comput. Assist. Surg.* 30.1 (Feb. 2025), p. 2466426. doi: 10.1080/24699322.2025.2466426

3. Z. Wang, **R. Butler**, J. J. van den Dobbelsteen, B. H. W. Hendriks, M. van der Elst, and J. Dauwels. "Towards Robust Object Detection in Unseen Catheterization Laboratories". In: *IEEE Int. Workshop Med. Meas. Appl.* IEEE, June 2024. doi: 10.1109/MeMeA60663.2024.10596906

2. J. Zeng, **R. Butler**, J. J. van den Dobbelsteen, B. H. W. Hendriks, M. van der Elst, and J. Dauwels. "Automatic Camera Pose Estimation by Key-Point Matching of Reference Objects". In: *Int. Conf. Acoust., Speech, Signal Process.* IEEE, June 2023. doi: 10.1109/ICASSP49357.2023.10095197

1. Y. Jiang, R. Dai, J. Zeng, **R. Butler**, T. Vijfvinkel, Y. Wang, J. J. van den Dobbelsteen, M. van der Elst, and J. Dauwels. "Object Detection and Person Tracking in CathLab with Automatically Calibrated Cameras". In: *Symp. Inf. Theory Signal Process. Benelux*. WIC, June 2022, p. 57. url: https://www.w-i-c.org/proceedings/proceedings_SITB2022.pdf