Towards Artificial Social Intelligence in the Wild

Sensing, Synthesizing, Modeling, and Perceiving Nonverbal Social Human Behavior

Raman, C.A.

DOI
10.4233/uuid:05fe4340-31bb-4c24-a827-69189aa2622b

Publication date
2023

Document Version
Final published version

Citation (APA)
Raman, C. A. (2023). *Towards Artificial Social Intelligence in the Wild: Sensing, Synthesizing, Modeling, and Perceiving Nonverbal Social Human Behavior*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:05fe4340-31bb-4c24-a827-69189aa2622b

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

TOWARDS ARTIFICIAL SOCIAL INTELLIGENCE IN THE WILD

Sensing, Synthesizing, Modeling, and Perceiving Nonverbal Social Human Behavior

Chirag Anantha Raman

Towards Artificial Social Intelligence in the Wild

Sensing, Synthesizing, Modeling, and Perceiving Nonverbal Social Human Behavior

Dissertation

for the purpose of attaining the degree of doctor
at the 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
Monday, 16th October 2023, at 10 : 00 a.m.

by

Chirag Anantha Raman

Master of Entertainment Technology,
Carnegie Mellon University, United States of America,
born on 25 September 1988 in Mumbai, India.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof. dr. ir. M.J.T. Reinders	Delft University of Technology, *promotor*
Prof. dr. M. Loog	Delft University of Technology, *promotor*
Dr. H. S. Hung	Delft University of Technology, *promotor*
Independent members:	
Prof. dr. P.S. César Garcia	Delft University of Technology
Dr. E. Gavves	University of Amsterdam
Dr. J. Tomczak	Eindhoven University of Technology
Dr. Z. Yumak	Utrecht University
Prof. dr. C. Jonker	Delft University of Technology, *reserve member*

Dissertation:
Towards Artificial Social Intelligence in the Wild: Sensing, Synthesizing, Modeling, and Perceiving Nonverbal Social Human Behavior

Department of Intelligent Systems
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
The Netherlands

Cover design by Javier Maldonado Acién | graftrei graftrei@gmail.com
Printed by Proefschriftspecialist | https://www.proefschriftspecialist.nl/

ISBN 978-94-93330-33-7

An electronic version of this dissertation is available at http://repository.tudelft.nl/

To my first teachers, Ma and Pa.

Contents

Summary **xi**

Samenvatting **xiii**

सारांश **xv**

સારાંશ **xvii**

1 Introduction **1**

1.1 Social Intelligence: Real and Artificial . . . 3
1.2 Methodological Approaches across Disciplines . . . 8
1.3 Moving Beyond the Laboratory: Considerations and Challenges in the Wild 10
1.4 Research Themes, Questions, & Contributions . . . 11
References . . . 15

I DATA ACQUISITION: SENSING & SYNTHESIS **23**

2 Curating Datasets of Social Interactions in the Wild: Data Collection for the Community by the Community **25**

Abstract . . . 26
2.1 Introduction . . . 26
2.2 Related Work. . . . 28
2.3 Data Acquisition . . . 30
2.4 Data Annotation . . . 32
2.5 Dataset Statistics . . . 34
2.6 Research Tasks . . . 35
2.7 Conclusion and Discussion . . . 38
References . . . 40
Appendices . . . 47
2.A Hosting, Licensing, and Organization . . . 47
2.B Datasheet For ConfLab . . . 49
2.C Sample Participant Report . . . 66
2.D Data Capture Setup Details. . . . 67
2.E Implementation Details . . . 68
2.F Additional Results . . . 69

2.G Reproducibility Checklist. 71

3 Synchronizing Multimodal Data at Acquisition for Capturing Social Interactions in the Wild 75

Abstract 76
3.1 Introduction 76
3.2 Related Work. 78
3.3 Our Approach 81
3.4 Experiments 85
3.5 Cost versus Latency Considerations 89
3.6 Conclusion 90
References 91

4 Synthesizing Training Data for Face-Related Tasks 95

Abstract 96
4.1 Introduction 96
4.2 Background: Synthesizing Faces 98
4.3 Related Work. 98
4.4 Synthethizing Expression-Based Wrinkles 100
4.5 Experiments and Results 104
4.6 Conclusion 107
References 108
Appendices 113
4.A Illustrating Tension Parameters 113
4.B Eye-Region Landmark Metrics 113
4.C Landmark Predictions on 300W 114
4.D Surface-Normals Predictions 114
4.E The 300W-winks Subset 117
4.F The Pexels Winks and Blinks Dataset. 117

II MODELING: FORECASTING & EXPLAINING SOCIAL BEHAVIOR 121

5 Adaptive Forecasting of Social Cues in Conversing Groups 123

Abstract 124
5.1 Introduction 124
5.2 Related Work. 126
5.3 Social Cue Forecasting: Task Formalization 127
5.4 Method Preliminaries 129
5.5 Social Processes: Methodology 130

5.6 Experiments and Results . . . 134
5.7 Discussion . . . 138
References . . . 139
Appendices . . . 146
5.A Detailed Results . . . 146
5.B Qualitative Visualizations . . . 150
5.C Additional Dataset Details . . . 152
5.D Implementation Details . . . 154
5.E Distinguishing Forecasting in Focused and Unfocused Interactions: A Meta Discussion . . . 155

6 Why Did This Model Forecast This Future? Information-Theoretic Saliency for Counterfactual Explanations of Probabilistic Regression Models 157
Abstract . . . 158
6.1 Introduction . . . 158
6.2 Related Work. . . . 160
6.3 Conceptual Grounding: Linking Saliency-Based Explanations to Counterfactual Reasoning . . . 161
6.4 Methodology: Closed-Form Saliency for Probabilistic Forecasting . . . 164
6.5 Experiments . . . 167
6.6 Conclusion . . . 172
6.7 Limitations and Potential Negative Societal Impact . . . 172
References . . . 173
Appendices . . . 178
6.A Broader Related Work: Explainable Methods for Time-Series Data Across Tasks and Domains. . . . 178
6.B Favorable Properties of Differential Entropy . . . 179
6.C Implementation Details for Experiments . . . 179
6.D Additional Results . . . 181
6.E Broader Discussion: Saliency & XAI with Domain Experts in the Loop . . . 182

III PERCEPTION: ANALYZING & QUANTIFYING SOCIAL PHENOMENA 185

7 Where is the Conversation? Investigating the Existence of Multiple Conversation Floors within an F-formation 187
Abstract . . . 188
7.1 Introduction . . . 188
7.2 Background . . . 190
7.3 Related Work. . . . 191

7.4 Methodology . . . 192
7.5 Dataset . . . 194
7.6 Experiments . . . 195
7.7 Conclusion . . . 199
References . . . 199

8 Perceived Conversation Quality in Spontaneous Interactions 203
Abstract . . . 204
8.1 Introduction . . . 204
8.2 Related work . . . 206
8.3 Perceived Conversation Quality . . . 208
8.4 Annotations, Validity, and Reliability . . . 211
8.5 Modeling Conversation Quality . . . 216
8.6 Results . . . 220
8.7 Discussion and Conclusion . . . 225
References . . . 227
Supplementary Material . . . 233
8.A PCQ Questionnaires . . . 233
8.B Feature Extraction Details . . . 234
8.C Class Imbalance Distribution . . . 238
8.D Additional Figures . . . 238

IV Discussion 239

9 Discussion 241
9.1 Curating Social Behavior Datasets . . . 242
9.2 Data Efficient and Adaptive Modeling of Social Behavior . . . 243
9.3 Synthesizing Social Human Behavior . . . 244
9.4 Ethics and Privacy: Behavior as Biometrics? . . . 245
9.5 Meta Discussion: Bringing Disciplines Closer - A Practitioner's Perspective 246
References . . . 250

Acknowledgments 253

Curriculum Vitæ 259

List of Publications 261

Summary

Over the last three decades, the social roots of human intelligence have come to influence the development of artificial intelligence (AI). Researchers in AI have moved beyond agents operating in isolation towards developing socially situated agents that can operate in the real world. Meanwhile, researchers in the social sciences have been leveraging AI techniques to analyze and theorize about social phenomena. Both these research endeavors came to be independently termed Artificial Social Intelligence (ASI), leading to the emergence of a field spanning several subdisciplines of the social and computational sciences.

This Thesis takes a holistic view of ASI and makes contributions toward both its historical goals. Moreover, the work presented here focuses on taking ASI research into natural real-world settings *in the wild*. The research is organized under three themes: *acquiring*, *modeling*, and *perceiving* social human behavior.

The Thesis begins by addressing the challenge of **data acquisition**. We propose a replicable data collection concept for curating datasets of real-world social human behavior, incorporating technical innovations and ethical considerations required for the noninvasive sensing of multimodal behavioral streams. To overcome the limited availability of real-world data, we also explore the potential of synthetic training data for downstream tasks.

Next, we tackle the challenge of **modeling** real-world social behavioral cues. Evidence from social psychology suggests that individuals uniquely adapt their behaviors to different conversation partners to sustain interactions. How can we jointly forecast these mutually dependent future cues of conversation partners? We propose a stochastic meta-learning method that adapts its forecasts to the unique dynamics of a conversation group given example behavior sequences. Thereby, it generalizes to unseen groups in a data-efficient manner by avoiding the need for group-specific models. Further, to facilitate the integration of data-driven and hypothesis-driven research, we propose a post hoc explanation framework for identifying timesteps that are salient to a forecasting model's predictions.

Finally, we contribute to a nuanced **perception** of social interactions by establishing evidence of multiple conversation floors within a single conversing group, in contrast to the prevailing implicit assumption in the automatic detection of conversation groups. We also develop an instrument for measuring the perceived quality of conversations at the individual and group levels.

Through these research themes, we provide novel contributions to the field of ASI, taking important steps toward the development of socially intelligent machines that can operate effectively in complex real-world settings.

Samenvatting

Gedurende de laatste drie decennia hebben de sociale aspecten van menselijke intelligentie een aanzienlijke invloed uitgeoefend op de ontwikkeling van artificiële intelligentie (AI). Onderzoekers binnen het domein van AI zijn afgestapt van het werken met geïsoleerde agenten en hebben zich toegelegd op de ontwikkeling van sociaal gesitueerde agenten die in de echte wereld kunnen functioneren. Tegelijkertijd hebben wetenschappers in de sociale wetenschappen AI-technieken toegepast voor de analyse van sociale fenomenen en het opstellen van theoretische modellen. Beide onderzoeksinspanningen werden onafhankelijk aangeduid als Artificial Social Intelligence (ASI), wat heeft geleid tot de opkomst van een interdisciplinair onderzoeksveld dat meerdere subdisciplines binnen de sociale en computationele wetenschappen omvat.

Dit proefschrift benadert ASI vanuit een holistisch perspectief, en levert bijdragen aan beide historische doelstellingen. Bovendien richt het hier gepresenteerde werk zich op het uitvoeren van ASI-onderzoek in natuurlijke, realistische situaties. Het onderzoek is gestructureerd rond drie thema's: het *verzamelen*, *modelleren* en *waarnemen* van sociaal menselijk gedrag.

Allereerst behandelt het proefschrift de uitdaging van het **verzamelen van gegevens**. Wij introduceren een herhaalbaar dataverzamelingsconcept om datasets van sociaal menselijk gedrag in realistische situaties samen te stellen, waarbij de technische innovaties en ethische aspecten worden geïntegreerd die noodzakelijk zijn voor het niet-invasief waarnemen van multimodale gedragspatronen. Om de beperkte beschikbaarheid van echte gegevens te ondervangen, onderzoeken we ook het potentieel van synthetische trainingsgegevens voor downstream taken.

Vervolgens richten we ons op het **modelleren** van realistische sociale gedragskenmerken. Studies in de sociale psychologie laten zien dat mensen hun gedrag specifiek aanpassen aan wie ze spreken, om een gesprek gaande te houden. Maar hoe kunnen we voorspellen hoe deze gedragingen zich in de toekomst zullen ontwikkelen, zeker als ze afhankelijk zijn van meerdere gesprekspartners? Wij introduceren een flexibele voorspellingsmethode, gebaseerd op stochastische meta-learning. Deze methode leert van voorbeeldgesprekken en past zich aan aan de unieke stroom van elk gesprek. Zo kan het model effectief worden toegepast op nieuwe, nog niet eerder geziene groepen, zonder dat er voor elke groep een apart model nodig is. Teneinde de integratie van datagedreven en hypothese-gedreven onderzoek te bevorderen, introduceren wij een post-hoc verklaringskader om tijdstappen te identificeren die relevant zijn voor de voorspellingen van een voorspellingsmodel.

Ten slotte dragen wij bij aan het genuanceerd **waarnemen** van sociale interacties door bewijs aan te voeren voor het bestaan van meerdere gespreksniveaus binnen één converserende groep. Hiermee betwisten we de gangbare impliciete veronderstelling in de automatische detectie van gespreksgroepen. Daarnaast ontwikkelen we een instrument om de waargenomen gesprekskwaliteit zowel op individueel als op groepsniveau te meten.

Door het verkennen van deze onderzoeksthema's dragen we niet alleen bij aan de vooruitgang van het vakgebied ASI, maar zetten we tevens cruciale stappen richting de ontwikkeling van sociaal intelligente machines die in staat zijn om effectief te functioneren in complexe omgevingen.

सारांश

पिछले तीस वर्षों में, मानव बुद्धि के सामाजिक आधारों ने कृत्रिम बुद्धिमत्ता (''आर्टिफिशियल इंटेलिजेंस (ए.आई.)") की प्रगति को प्रभावित करने में महत्वपूर्ण भूमिका निभाई है। ए.आई. शोधकर्ता एकांत में कार्यरत ''एजेंटों" से आगे बढ़कर ऐसे पारस्परिक संवादात्मक एजेंटों के विकास में प्रयासित हैं जो वास्तविक दुनिया में कार्य करने योग्य हों। इस दौरान, सामाजिक विज्ञान के शोधकर्ता सामाजिक घटनाओं से संबंधित सिद्धांतों का अध्ययन और निर्माण करने के लिए ए.आई. विधियों का उपयोग कर रहे हैं। ये दोनों शोध प्रयास स्वतंत्र रूप से कृत्रिम सामाजिक बुद्धिमत्ता (''आर्टिफिशियल सोशियल इंटेलिजेंस (ए.एस.आई.)") के रूप में जाने गये, जिसके परिणामस्वरूप एक ऐसे क्षेत्र की स्थापना हुई जिसमें सामाजिक और संगणना (कम्प्यूटेशनल) विज्ञान की विभिन्न शाखाएं शामिल हैं।

यह शोध-प्रबंध (थीसिस) ए.एस.आई. का समग्र दृष्टिकोण अपनाती है और इसके दोनों ऐतिहासिक लक्ष्यों की दिशा में योगदान देती है। विशेष रूप से, यहां प्रस्तुत शोध नियंत्रित वातावरण से परे वास्तविक दुनिया के परिदृश्यों में ए.एस.आई. के अनुप्रयोग की खोज पर केंद्रित है। यह जांच तीन मुख्य विषयों पर संरचित है: सामाजिक मानव व्यवहार से संबंधित *दत्त-सामग्री/''डेटा" संग्रह, प्रतिमान/''मॉडल" विकास,* और व्यवहार के संज्ञानात्मक पहलुओं की *अनुभूति*।

यह शोध-प्रबंध **डेटा संग्रह** की चुनौती को संबोधित करते हुए आरम्भ होती है। हम वास्तविक दुनिया के सामाजिक मानव व्यवहार के डेटासेट एकत्र करने के लिए एक प्रतिकृति डेटा संग्रह अवधारणा को प्रस्तावित करते हैं। विशेष रूप से, हम विभिन्न व्यवहार धाराओं की गैर-आक्रामक संवेदन के लिए आवश्यक तकनीकी नवाचारों और नैतिक विचारों को शामिल करते हैं। इसके अतिरिक्त, हम वास्तविक दुनिया के डेटा की कमी की भरपाई के साधन के रूप में एआई कार्यों के लिए कृत्रिम रूप से उत्पन्न प्रशिक्षण डेटा का उपयोग करने की संभावना की भी जांच करते हैं।

तदुपरांत, हम वास्तविक दुनिया के सामाजिक व्यवहार संबंधी संकेतों के **मॉडलिंग** की चुनौती से निपटते हैं। सामाजिक मनोविज्ञान के प्रमाण से पता चलता है कि व्यक्ति बातचीत को बनाए रखने के लिए अलग-अलग वार्तालाप भागीदारों के साथ अपने व्यवहार को विशिष्ट रूप से अनुकूलित करते हैं। हम वार्तालाप साझेदारों के इन परस्पर निर्भर भविष्य के संकेतों का संयुक्त रूप से पूर्वानुमान कैसे लगा सकते हैं? हम एक ''स्टोकेस्टिक, मेटा-लर्निंग" पद्धति का प्रस्ताव प्रस्तुत करते हैं जो अपने पूर्वानुमानों को उदाहरण व्यवहार अनुक्रम दिए गए वार्तालाप समूह की अद्वितीय गतिशीलता के अनुरूप बनाती है। इस प्रकार, यह समूह-विशिष्ट मॉडलों की आवश्यकता से बचकर डेटा-कुशल तरीके से अनदेखे समूहों का सामान्यीकरण करता है। इसके अलावा, डेटा-संचालित और परिकल्पना-संचालित अनुसंधान के एकीकरण को सुविधाजनक बनाने के लिए, हम पूर्वानुमान मॉडल की भविष्यवाणियों के लिए मुख्य समय-चरणों की पहचान करने के लिए एक ''पोस्ट हॉक" स्पष्टीकरण साध्य का प्रस्ताव प्रस्तुत करते हैं।

अंत में, हम वार्तालाप समूहों की स्वचालित पहचान में प्रचलित अंतर्निहित धारणा के विपरीत, एक ही समूह के अंतर्गत एकाधिक समांतर वार्तालाप के अस्तित्व का प्रमाण स्थापित करके सामाजिक बातचीत की एक सूक्ष्म **अनुभूति** में योगदान देते हैं। आगे, हम व्यक्तिगत और समूह स्तर पर बातचीत की कथित गुणवत्ता को मापने के लिए एक उपकरण भी विकसित करते हैं।

इन शोध विषयों के माध्यम से, हम ए.एस.आई. के क्षेत्र में नवीन योगदान प्रदान करते हुए, ऐसे सामाजिक रूप से बुद्धिमान मशीनों के विकास की दिशा में महत्वपूर्ण कदम उठाते हैं जो वास्तविक दुनिया के जटिल परिदृश्यों में कुशलता से काम कर सकते हैं।

સારાંશ

છેલ્લા ત્રણ દાયકાઓમાં, માનવ બુદ્ધિના સામાજિક મૂળે કૃત્રિમ બુદ્ધિ ("આર્ટિફિશિયલ ઇન્ટેલિજન્સ (એ.આઇ.)") ના વિકાસને પ્રભાવિત કરવામાં મહત્વપૂર્ણ ભૂમિકા ભજવી છે. એ.આઇ. સંશોધકો એકલતામાં કાર્યરત "એજન્ટો"થી આગળ વધીને એવાં પરસ્પર સંવાદાત્મક એજન્ટો વિકસાવવા તરફ અગ્રસર થયા છે જે વાસ્તવિક દુનિયામાં કાર્ય કરી શકે છે. આ દરમિયાન, સામાજિક વિજ્ઞાનના સંશોધકો સામાજિક ઘટનાઓ વિશેના સિદ્ધાંતોનું વિશ્લેષણ અને નિર્માણ કરવા માટે એ.આઇ. તકનીકોનો ઉપયોગ કરી રહ્યાં છે. આ બંને સંશોધન પ્રયાસો સ્વતંત્ર રીતે કૃત્રિમ સામાજિક બુદ્ધિ ("આર્ટિફિશિયલ સોશિયલ ઇન્ટેલિજન્સ (એ.એસ.આઇ.)") ના નામથી ઓળખાવામાં આવ્યા, જે સામાજિક અને ગણના (કોમ્પ્યુટેશનલ) વિજ્ઞાનની અનેક પેટાશાખાઓમાં ફેલાયેલું ક્ષેત્ર છે.

આ શોધ-પ્રબંધ (થીસીસ) એ.એસ.આઇ. નો સર્વગ્રાહી દૃષ્ટિકોણ લે છે અને તેના બંને ઐતિહાસિક ધ્યેયો તરફ યોગદાન આપે છે. તદુપરાંત, અહીં પ્રસ્તુત કાર્ય એ.એસ.આઇ. સંશોધનને કુદરતી વાસ્તવિક-વિશ્વની પરિસ્થિતિઓમાં લેવા પર કેન્દ્રિત છે. આ સંશોધન ત્રણ વિષયો હેઠળ આયોજિત છે: સામાજિક માનવ વર્તનની *માહિતી/"ડેટા" સંગ્રહ*, *પ્રતિકૃતિ/"મોડેલિંગ"*, અને *અનુભૂતિ*.

આ શોધ-પ્રબંધ (થીસીસ) **ડેટા સંગ્રહ** ના પડકારને સંબોધીને શરૂ થાય છે. અહીં વાસ્તવિક-વિશ્વના સામાજિક માનવ વર્તણૂકના "ડેટાસેટ્સ" એકત્રિત કરવા માટે એક પુનરુત્પાદન યોગ્ય ડેટા સંગ્રહ ના ખ્યાલનો પ્રસ્તાવ અમે આપીએ છીએ. ખાસ કરીને, અહીં અમે વિભિન્ન વર્તન પ્રવાહની બિન-આક્રમક સંવેદના માટે જરૂરી તકનીકી નવીનતાઓ અને નૈતિક વિચારણાઓનો સમાવેશ કરીએ છીએ. વધુમાં, વાસ્તવિક દુનિયાના ડેટાની મર્યાદિત ઉપલબ્ધતાને દૂર કરવા માટે, અમે એ.આઇ. કાર્યો માટે કૃત્રિમ તાલીમ ડેટાની ક્ષમતા પર અન્વેષણ કરીએ છીએ.

તદુપરાંત, અમે વાસ્તવિક-વિશ્વના સામાજિક વર્તણૂકીય સંકેતોના **મોડેલિંગ** પડકારનો સામનો કરીએ છીએ. સામાજિક મનોવિજ્ઞાનના પુરાવા સૂચવે છે કે દરેક વ્યક્તિ ક્રિયાપ્રતિક્રિયાને ટકાવી રાખવા માટે તેમના વર્તનને વાતચીતના વિવિધ ભાગીદારો સાથે અનન્ય રીતે અનુકૂલિત કરે છે. વાર્તાલાપ ભાગીદારોના આ પરસ્પર નિર્ભર ભાવિ સંકેતોની આપડે સંયુક્ત રીતે આગાહી કેવી રીતે કરી શકીએ? આ વિષયમાં અમે "સ્ટોકેસ્ટિક મેટા-લર્નિંગ" પદ્ધતિનો પ્રસ્તાવ આપીએ છીએ જે તેના અનુમાનને વાતચીત જૂથની અનન્ય ગતિશીલતા સાથે અનુકૂલિત કરે છે, ઉદાહરણ તરીકે વર્તન ક્રમ. આમ, તે જૂથ-વિશિષ્ટ મોડલ્સની જરૂરિયાતને ટાળીને ડેટા-કાર્યક્ષમ રીતે અદ્રશ્ય જૂથોને સામાન્ય બનાવે છે. વધુમાં, ડેટા-આધારિત અને પૂર્વધારણા-સંચાલિત સંશોધનના એકીકરણને સરળ બનાવવા માટે, અમે આગાહી મોડલના અનુમાનો માટે મહત્ત્વપૂર્ણ હોય તેવા સમય પગલાંને ઓળખવા માટે "પોસ્ટ-હોક" સમજૂતી માળખું પ્રસ્તાવિત કરીએ છીએ.

અંતમાં, અમે વાતચીત જૂથોની સ્વચાલિત શોધમાં પ્રચલિત ગર્ભિત ધારણાથી વિપરીત, એક જ વાર્તાલાપ જૂથની અંદર અનેક સમાંતર વાર્તાલાપના અસ્તિત્વના પુરાવા સ્થાપિત કરીને સામાજિક ક્રિયાપ્રતિક્રિયાઓના સૂક્ષ્મ **અનુભૂતિ** માં યોગદાન આપીએ છીએ. આગળ, અમે વ્યક્તિગત અને જૂથ સ્તરે વાતચીતની કથિત ગુણવત્તાને માપવા માટે એક સાધન પણ વિકસાવીએ છીએ.

આ સંશોધન વિષયો દ્વારા, અમે એ.એસ.આઇ. ના ક્ષેત્રમાં નવતર યોગદાન પ્રદાન કરીને આવા સામાજિક રીતે બુદ્ધિશાળી મશીનોના વિકાસ તરફ મહત્વપૂર્ણ પગલાં લઇએ છીએ જે વાસ્તવિક-વિશ્વના જટિલ સંજોગોમાં અસરકારક રીતે કાર્ય કરી શકે છે.

1

Introduction

So once you do know what the question actually is,
you'll know what the answer means

— Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

1

Social human interaction is a spectacularly elaborate process. Information born of a medley of mental and physical processes involving our desires, beliefs, feelings, and intentions is manifested through a complex interplay between verbal and nonverbal messages. These are then transmitted over often imperfect and noisy channels in the hope that when they are decoded by perceivers, the original meaning is preserved. Yet, social interactions and relationships are critical to our daily lives and wellbeing [1]. In fact, so central is the social context to our existence that researchers have long hypothesized that human intelligence evolved primarily to adapt to the complexities of social life [2, 3].

The influence of these social roots of human intelligence on artificial intelligence (AI) research gained impetus in the 1990s. In this era, two distinct research endeavors emerged independently and were subsequently referred to as Artificial Social Intelligence (ASI): one rooted in the computational sciences, focusing on developing autonomous agents, and the other originating from the social sciences, utilizing computational tools to understand social human behavior. Recognizing the limitations of studying isolated agents, AI researchers began accounting for the social and cultural environment to develop *socially situated* agents capable of operating in the real world [4–7]. As researchers delved into the development of machines that could interact with their surroundings and humans, it became apparent that these artificial intelligence agents also required artificial *social* intelligence. Simultaneously, sociologists began applying AI techniques to analyze social phenomena and construct theories pertaining to human behavior [8, 9]. This intersection of AI techniques and the study of social behaviors also came to be known as the field of Artificial Social Intelligence.

Over the years, these motivations have intertwined to crystallize into two closely related interdisciplinary domains: affective computing and social signal processing [10]. Affective computing places human emotion at the core, exploring how affective factors influence interactions between humans and technology, and how sensing and generating affect can inform our understanding of human behavior [10, Chap. 2]. On the other hand, the broader field of social signal processing aims to model and comprehend the social meaning of nonverbal human behavior in interactive contexts from a machine perspective [10, Chap. 7]. However, despite significant progress in these fields, a central drawback persists. The majority of research has primarily occurred in controlled laboratory settings or semi-controlled pre-arranged interactions [11, 12]. While such settings afford researchers the advantage of isolating specific phenomena of interest for study, a crucial question arises: to what extent do findings from controlled settings reflect the realities of uncontrolled, real-world settings?

The objective of this Thesis is to bring ASI research directly into real-world settings, contributing to both historical motivations: advancing the development of socially intelligent machines and assisting domain experts in the social sciences to gain new insights into social human behavior. The research presented here specifically focuses on real-world

interaction settings *in the wild*. This work adopts an inherently interdisciplinary outlook, encompassing machine learning, computer vision/graphics, affective computing, social signal processing, and distributed systems. Consequently, an implicit goal of this work is to bridge the divide that exists between these disciplines, which arises in part due to the differences in methodologies and focus areas, resulting in contrasting values and goals. Researchers from these distinct fields might find familiarity in the data-driven and hypothesis-driven methodologies employed in this work and appreciate how these methods complement and support each other.

This introductory chapter is organized as follows. I begin by reviewing various attempts at defining human social intelligence and its artificial counterpart in Section 1.1. Specifically, to emphasize the bidirectional exchange between AI and social sciences, I propose three broad goals of ASI. In Section 1.2, I review the broad methodological commonalities and contrasts between disciplines. I then discuss the challenges involved in taking human behavior research into real-life settings in Section 1.3. Finally, I conclude by describing the specific research themes and questions considered in this Thesis and summarizing concrete contributions it makes in Section 1.4.

1.1 Social Intelligence: Real and Artificial

1.1.1 The Social Roots of Intelligence

Defining Social Intelligence. Social intelligence has been explicitly viewed as an integral aspect of human intelligence since as early as 1920, when Thorndike [13] distinguished social from mechanical and abstract intelligence. He defined social intelligence as "the ability to understand and manage men and women, boys and girls—to act wisely in human relations". While Thorndike did not build any theory of social intelligence, this simple definition encapsulated the two primary components that would guide future definitions over the next century [14]: the cognitive (understanding others) and the behavioral (acting effectively in social situations). Nevertheless, defining social intelligence remains a difficult task. One issue is the lack of consensus [15, 16]: some researchers emphasize the cognitive aspects [17], others the behavioral [18], while yet others focus on a psychometric foundation in terms of "the ability to perform well on tests that measure social skills" [19]. Another issue is that socially intelligent behavior is contextual. Strang [20] argued that social intelligence is not a "unit characteristic but rather a complex pattern of behavior", and may vary for the same individual depending on the interaction partners, situation, and time. Finally, researchers have long questioned whether social intelligence is a distinct construct at all. In 1930 Strang [20] posited that general and social intelligence may be inextricably linked. Based on correlations between scores on social and general intelligence tests, she argued that the two may be unanalyzable parts of "a total organic attitude, involving attitudes of

mind, emotional conditions, ingrained habits and conditioned behavior" [21]. Confounded by concerns over the validity and reliability of the measures [20], this perspective persisted for decades owing to the empirical difficulty in separating social intelligence from other related constructs such as academic intelligence [19]. More sophisticated recent designs have now established evidence for the distinguishing social from academic intelligence [18] as well as the cognitive and behavioral components of social intelligence itself [22]. Despite such progress, the dimensions and measures of social intelligence still vary significantly across studies, and a unified definition remains to be established.

An Evolutionary Perspective. In the absence of a clear definition and measure, recent stances follow an evolutionary perspective; here social intelligence is viewed as the manifestation of the *theory of mind* (ToM) [23] ToM refers to the ability to ascribe mental states such as desires, beliefs, feelings and intentions to oneself and others [24–26]. Knowing what people want, think, feel, and intend enables one to interpret their behavior and make predictions about how they will act [25, 27]. Moreover, the absence of ToM may relate to an impairment in social communication and high-level control of actions, as found to be the case in individuals with autism spectrum disorder and Asperger syndrome [25, 26, 28].

The theory of mind can trace its roots to theories from cognitive evolution that have come to be collectively called the *social intelligence hypothesis* [29–31]. The hypothesis suggests that our higher intellectual faculties do not merely help navigate social situations as a consequence, but may have primarily evolved to adapt to the complexities of social living. In his seminal work on *the social function of intellect*, Humphrey [31] argued that more than the physical daily problems confronting primates (apes and humans), such as finding and extracting food, it is the competitive social maneuvering—the ability to recognize individuals, track relationships and deceive one another—that has driven the development of our sophisticated intelligence and large brains. In particular, it was Humphrey's emphasis on this anticipation, counter-anticipation, and manipulation of behaviors and minds of others that led to the ToM becoming a research focus in comparative and developmental psychology [32]. The Social Intelligence Hypothesis complements theories surrounding the importance of social context in developing human intelligence proposed as early as the 1920s: Lev Vygotsky's theory of cognitive development emphasizes that individual intelligence emerges as a result of biological factors (embodiment) that interacts with a physical, and especially, a social environment (social situatedness) through a developmental process [7][1].

[1]Unfortunately, Vygotsky's work from the 1920s and 1930s took a while to widely influence research, partly because it only reached the Western world in the 1960s, with the first public translation appearing in 1962. This in turn may be because it was banned in the Soviet Union from the mid-1930s to the mid-1950s. See [7, Sec. 2] for a detailed discussion of his ideas.

1.1.2 Artificial Social Intelligence

How do these theories about human intelligence and its development influence the advancement of artificial intelligence? Over the last three decades, there have been distinct views on the matter, several of which have used the term *Artificial Social Intelligence* (ASI) to encapsulate research goals and methods. The earliest explicit use of the term was in 1994, when Bainbridge et al. [8] used it to discuss the use of AI techniques within Sociology, the discipline focusing on topics pertaining to social structure, social class, and social institutions in society at large. Independently, inspired by the social intelligence hypothesis, Dautenhahn [6] called for a field with the name of ASI towards developing interactive autonomous robots that could lead to individualized robot societies. However, as Kappas et al. [33] observe, the term ASI did not catch on. Nevertheless, much of what is currently explicitly referred to as ASI lies within the context of affective computing, social robotics, and social signal processing. Figure 1.1a illustrates the overlap between subdisciplines relevant to ASI research. Kappas et al. [33] explicitly refer to the field of ASI, mentioning only affective computing and social robotics. While the authors acknowledge that the terms in Figure 1.1a are not mutually interchangeable, the contemporary use of *ASI* is largely restricted to the development of interactive systems that understand their social context and interaction partners: ASI in the sense of an artificial *agent with* social intelligence. Little, comparatively, is discussed in current literature about the field of artificial social intelligence as the primary focus, especially what lies within its research scope outside the realm of affective computing and social robotics.

To take a broader perspective, in this subsection, I propose three goals of ASI drawing upon distinct historical motivations spanning disciplines (see Figure 1.1b). These are: (i) developing socially-aware AI systems; (ii) AI-assisted social theory building; and (iii) socially situated development of AI. Rather than directly defining the term ASI, I argue that it is of more practical value to describe the various goals, problems, and methods that fall within its scope. The primary motivation in broadening this scope beyond its conventional use is simple: ASI ought to subsume the *bidirectional* reciprocity between artificial intelligence and what we know of social intelligence. The conventional use emphasizes a unidirectional transfer of knowledge from the social sciences to AI for building an applied system with social intelligence. A bidirectional perspective also encapsulates the use of AI techniques for advancing social theory, independent of whether an interactive agent is involved in the process. Moreover, the view of ASI presented here also incorporates foundational AI research that draws inspiration from how natural intelligence develops within a social context. Beyond this conceptual contribution, this Thesis makes novel contributions to the first two of the three proposed goals.

1

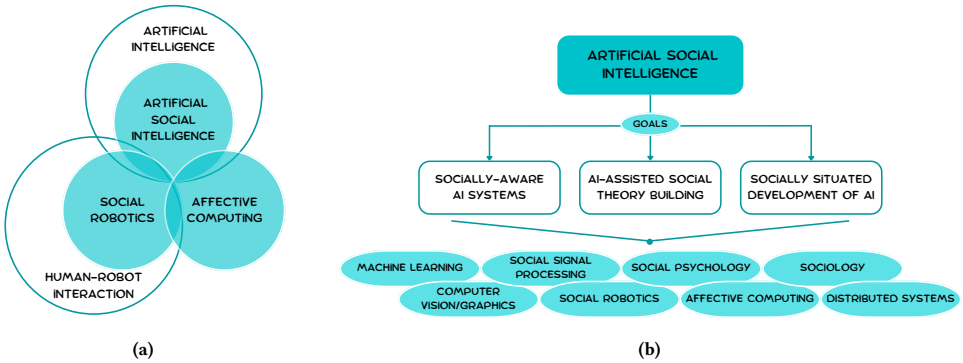


Figure 1.1: Illustrating the interconnections among subdisciplines relevant to Artificial Social Intelligence (ASI) research. (a) Kappas et al. [33] explicitly characterize ASI as a field, suggesting that much of its content can be found within the domains of Affective Computing and Social Robotics. The primary motivation is developing artificial agents *with* social intelligence; **(b)** In this work, I adopt a distinct view of ASI, highlighting the proposed objectives and the diverse disciplines they encompass. This perspective emphasizes bidirectional reciprocity between the computational and social fields, expanding the previous notion of ASI with the original dual motivations: (i) *ASI* as in the AI-assisted understanding of social intelligence; and (ii) utilizing insights from social intelligence for training artificial intelligence.

Developing Socially-Aware AI Systems. Since humans are fundamentally social, a long-standing research goal has been to endow interactive artificial agents with human-inspired social intelligence [6, 34, 35]. In order to effectively communicate with people, such systems need to participate in a bidirectional exchange of social meaning. This high-order semantic meaning—the attitudes, intents, feelings, mental states, personalities, etc. of people—constitutes a *social signal* [36] that is embodied through people's behavior and transferred through a set of low-level *behavioral cues*. Here, the term behavioral cue describes a set of short observable temporal changes in physical or physiological activity, with examples including gaze exchanges, smiles, head nods, and winks [36, 37]. For communicating with humans, interactive agents need to be able to *sense* such cues, *perceive* the social signal embedded within, and *synthesize* interpretable cues to convey their social information. Even when a system is not designed to interact with humans directly, the sensing and perception components are critical for social awareness. More recently, in an effort to go beyond the limitations of modeling only external behavior while ignoring internal mental states, researchers have leveraged cognitive models. Specifically, given its integral role in social intelligence, researchers have argued that endowing agents with an artificial theory of mind is crucial to the agents' abilities in operating alongside humans and multiagent systems [23, 38].

AI-Assisted Social Theory Building. One of the earliest explicit definitions of Artificial Social Intelligence dates back to 1994, when Bainbridge et al. [8] used it to describe research

reciprocity between AI and Sociology:

> "Broadly defined, Artificial Social Intelligence (ASI) is the application of machine intelligence techniques to social phenomena. ASI includes both theory building and data analysis." [8]

The discussion reflected the growing connections between the two disciplines at the time, owing to "a growing interest among researchers in artificial intelligence in the socially situated agent, and a growing interest among sociologists in using artificial intelligence techniques for theorizing about social phenomena" [9]. Here, special emphasis was laid on the use of data- and theory- driven computational simulations towards rendering theories more rigorous, connecting scattered hypotheses into coherent theoretical frameworks, discovering hidden assumptions, or inspiring new theories altogether [8]. Since then, AI has seen much progress, and there are some recent examples of simulations being used to test social theories [39]. Nevertheless, research within AI and social theory largely remain independent endeavors. Recently, in order to operationalize AI-driven social theory beyond relying on simulations, Mökander and Schroeder [40] proposed three essential requirements for AI systems: semanticization, transferability, and generativity. Similarly, to aid researchers at the intersection of machine learning and the social sciences, Radford and Joseph [41] outlined a *theory in, theory out* framework, outlining how social theory can help build machine learning models (theory in), and a checklist of the potential uses of the model (theory out).

Socially Situated Development of AI. Evolutionary speaking, given that the social context is crucial for the development of natural intelligence, could a social (and cultural) embedding similarly aid artificial intelligence? Motivated by this question, since the 1990s and early 2000s, AI researchers have argued for the development of AI systems by embedding them within a social environment where it learns by interacting with humans [6, 42]. Known as *Socially Situated AI* [7, 43], this argument for the simultaneous development of technical and social intelligence is complemented by a cognitive developmental perspective: children are born in to a social environment and grow up as social beings alongside acquiring technical skills that are required for specific tasks [34]. While traditional AI research between the 1950s and 1980s paid little attention to social factors and learning/development—Gardner [44] argued that accounting for the "murky concepts" of affect, context, and cultural factors would confound finding the "essence" of human cognition—the developmental underpinnings of Socially Situated AI are also found in Turing's works. In his 1950 paper *Computing Machinery and Intelligence* [45], Turing discusses the so-called *child machines*: instead of simulating the adult mind, the idea is to produce a machine simulating that of a child's, whose education "could follow the normal teaching

of a child". Noting that the Turing Test is a test of human social intelligence rather than of a putative *general intelligence* [46], some researchers have also argued for the evaluation of AI in the social context. Edmonds [47] postulated that in order to pass the Turing Test over any period of extended time, it is necessary to embed the AI entity into society.

Recently, within the past two years, the idea of agents learning through social interactions has received renewed attention [48, 49]. Krishna et al. [48] also proposed the framework of socially situated AI [48], without acknowledging any of the past works using the same term. In contrast to the prior conceptualizations of the framework however, they formalize the task of socially situated learning as a reinforcement learning problem. Elsewhere, Bolotta and Dumas [49] propose the framework of *Social Neuro-AI* developing three research axes towards aligning interactions between natural and artificial intelligence: biological plausibility, temporal dynamics, and social embodiment. Specifically, they also identify multi-agent reinforcement learning (MARL) and active inference as promising tools towards social learning.

1.2 Methodological Approaches across Disciplines

The aforementioned goals, in as much as I believe they ought to fall under the purview of ASI, illustrate that ASI research has implications for several disciplines. Consequently, the pursuit of these goals requires a broad understanding of the commonalities and contrasts between the methodologies native to these disciplines. To characterize how this Thesis makes contributions towards the first two of the goals of ASI, I next provide a categorization of these approaches. This categorization is not meant as a comprehensive taxonomy. Rather, its purpose is to serve as a window into understanding the often contrasting goals of different disciplines, and consequently, the values and assumptions guiding the people engaging in the enterprise of science, to the extent that they are aware of these.

Top-Down and Bottom-Up Approaches. At the broadest categorization, research in ASI has leveraged two antipodal strategies from AI research that are termed *top-down* and *bottom-up* approaches. The earliest foundations of these concepts were laid by Alan Turing in his 1948 manifesto, where he contrasted machines built for a definite purpose from those constructed from some kind of standard components [50]. Historically, top-down approaches dealt with "high-level symbolic processes that reflect the complex thought processes of which humans are capable" [8, p. 409]. Symbolic process models conceptualize the world using high-level human-readable representations of problems, where procedures (logical rules) and words are used to describe behaviors and actions. In contrast, bottom-up approaches such as neural networks are predominantly numeric frameworks, concerned with modeling low-level processes such as the functioning of a bundle of neurons, "with the hope that eventually they could work their way up to the level of human consciousness"

[8, p. 409]. The bottom-up approach is often referred to as *connectionism* given its assertion that intelligence arises not in the manipulation of symbols but in the connections between neurons. Over the decades, other similar terms in different domains have come to describe the broad conceptual underpinnings of these strategies. Trading precision for comprehension, some analogous terms capturing the contrasting focus areas of top-down vs bottom-up approaches are *confirmatory* vs. *exploratory*, *hypothesis-driven* vs. *data-driven*, and *inductive* vs. *deductive*. Note that this brief overview is not meant to do full justice to such a complex topic, but to only provide you, the reader, with a representative picture of the historical roots of these contrasting strategies.

While the terms *top-down* and *bottom-up* are overloaded and allude to different concepts across domains [51–58], in this Thesis they incorporate the primary goal of the research and where it begins in addition to the specific methodology used. When the primary purpose is to gain semantic insights into specific behavioral phenomena, the resulting approach is referred to as top-down. In contrast, the primary focus of bottom-up strategies is on modeling patterns in the available data. Consequently, there exists a semantic gap[2] in the expected human-understandable insights the research strategy provides into social phenomena. Methodologically, hypothesis- and data-driven research designs correspond more naturally to top-down and bottom-up strategies respectively. However, it is entirely possible, and indeed commonplace, for top-down approaches to use techniques that would qualify as data-driven. For instance, while exploratory factor analysis [60] is a data-driven technique, it is employed to understand human-interpretable relationships between input dimensions in the data. Consequently, the incorporation of the focus areas beyond the methodology constitutes a nuance in the way this Thesis refers to top-down and bottom-up approaches.

Methodologies and Disciplines. A holistic consideration of how we may endow machines with social intelligence spans disciplines; these include machine learning, computer vision/graphics, social psychology, affective computing, social signal processing, and distributed systems. How do the research strategies map to research disciplines? It is perhaps unsurprising that top-down methods are common in the disciplines of social psychology and sociology, while bottom-up approaches dominate machine learning and its cousins. Research in the social fields often begins with a hunch, or a belief about the world, that the researcher wishes to test [11, Ch. 2]. These hypotheses may result from the researcher's dissatisfaction with previous theories and explanations of phenomena, or from their own

[2] The term *semantic gap* was originally defined within an information retrieval setting as "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" [59]. Its use here is beyond the strict scope of its original definition to capture the sense of a difference in semantic information a researcher may obtain about a specific social phenomenon from contrasting research strategies.

personal experiences. Nevertheless, the overarching goal is to expand our understanding of social influence at the level of individuals—in the case of social psychology—or groups, institutions, and even societies at large—as is the case in sociology. In contrast, research in machine learning, even when dealing with tasks surrounding behavioral data, is primarily concerned with finding patterns in available data. Primary goals notwithstanding, several analogs exist in the specific methods employed across these domains. Revisiting the example of exploratory factor analysis, while the tool is most commonly used by psychologists, it is one of several unsupervised learning techniques. The curious reader may further refer to articles by Allen et al. [61], Liem et al. [62], and Nilsen et al. [63] for a deeper discussion on the commonalities and contrasts between the domains and methodologies.

1.3 Moving Beyond the Laboratory: Considerations and Challenges in the Wild

A great deal of research into social human behavior occurs in laboratory settings. This is because it enables the researcher to orchestrate events so that extraneous factors do not influence the phenomena under study. This is typically the case when the researcher is interested in studying the effect of varying some variable, called the independent variable, on an outcome of interest, namely the dependent variable. Here, the key to a good experiment is maximizing *internal validity* [11, p. 36]: ensuring that nothing other than the independent variable affects the dependent variable.

However, this experimental control in the laboratory often comes at the cost of realism. One way to increase realism is to conduct experiments in the real world, but the natural setting makes it harder to control for extraneous variables. This gives rise to concerns over *external* or *ecological validity*[3] [11, p. 37]: to what extent do the results of an experiment generalize to other situations and people? In social psychology, this trade-off between internal and external validity is called *the basic dilemma of the social psychologist* [66].

Today, the phrase *in the wild* has come to broadly describe research that seeks to understand or operate in naturalistic settings from everyday living. The term is widely used in several disciplines: in human-computer interaction to evaluate technology interventions and account for user experiences in everyday lives [67]; in deep learning to refer to the gap between cutting-edge research and its applications in practice [68]; in computer vision to motivate the need for moving beyond restricted supervisory labels towards open-set/domain visual recognition and task-level transfer [69]; and in affective computing to evaluate emotion recognition methods in noisy real-world conditions [70].

[3]Readers interested in the evolution of research terminology may enjoy essays by Hammond [64] and Kihlstrom [65] on the matter of *ecological validity*.

Beyond the trade-off between internal and ecological validity, researching social human behavior in the wild poses several other practical challenges:

- **Noninvasive and Distributed Sensing**: How can we design a noninvasive sensor setup to avoid invalidating the naturalness of the behavior? When the study involves several participants, how can we deploy multisensor setups such that the collected data is well synchronized across sensor streams and modalities?
- **Data Fidelity vs. Privacy Preservation**: Capturing high-fidelity data often bears the risk of revealing sensitive participant information. For instance, high-frequency speech facilitates the extraction of verbal content of real interactions. Similarly, high-resolution video makes it easier to reveal participant identities and, possibly, lip movements, especially when faces are captured. How do we select behavioral streams that capture social dynamics while protecting participant privacy?
- **Ethical Considerations**: How do we design experiments to respect participant consent and provide them with agency over their data? How can we comply with standards of responsible data collection and sharing?
- **Scene and Data Noise**: Real interactions can evolve in unpredictable ways making it challenging to obtain clean and usable data, especially in complex conversational scenes [71]. With video, occlusions, unfavorable lighting, and failure to track individuals beyond camera coverage constitute possible challenges. With audio, ambient noise and cross-talk between individual microphones make it hard to isolate the primary source of the speech data.

1.4 Research Themes, Questions, & Contributions

Given the interdisciplinary nature of its subject matter, the contrasting approaches across domains, and the various challenges characterizing the in-the-wild setting, this Thesis is organized under three research themes: i. *data acquisition*: sensing and synthesis for supporting downstream bottom-up modeling and top-down analyses; ii. *modeling*: data-efficient methods for predicting real-world social behavior and obtaining post hoc data-driven insights; and iii. *perception*: hypothesis-driven analysis and development of instruments for quantifying relevant social phenomena. Figure 1.2 illustrates the contributions this Thesis makes in supporting diverse workflows surrounding ASI research.

1.4.1 Data Acquisition: Sensing & Synthesis

If we are to endow machines with the ability to perceive and operate in the real world, common wisdom entails that we need to obtain ecologically valid data from the real world. The previously discussed trade-offs between data fidelity, participant privacy, and ethical considerations confound data acquisition efforts beyond the laboratory setting. How can

1

we then record the dynamics of unscripted human interactions in the real world? The highly instrumented wired setups common in lab studies are impractical for sensing natural free-standing interactions. How can we then design wireless multisensor and multimodal sensing setups to preserve ecological validity while also allowing for fine-grained analysis of social phenomena?

Chapter 2 follows a *dataset by the community for the community* ethos and proposes a data collection concept viewing conferences as living labs called ConfLab. The chapter also describes the first instantiation of this concept at a major international conference and the resulting dataset and benchmark. Crucially, we discuss the often overlooked and underappreciated aspects of conducting an in-the-wild dataset collection: participatory design, engineering innovations, responsible data sharing, and the design choices that empower participant agency. In doing so, the broader goal is to serve as a template such that data collection efforts may be replicated by the community, diminishing the logistical burden on any single research group.

Chapter 3 details the technical innovation surrounding data synchronization in ConfLab. Specifically, we propose a modular solution for synchronizing wireless sensors across modalities at acquisition itself for in-the-wild behavior research. Traditional approaches

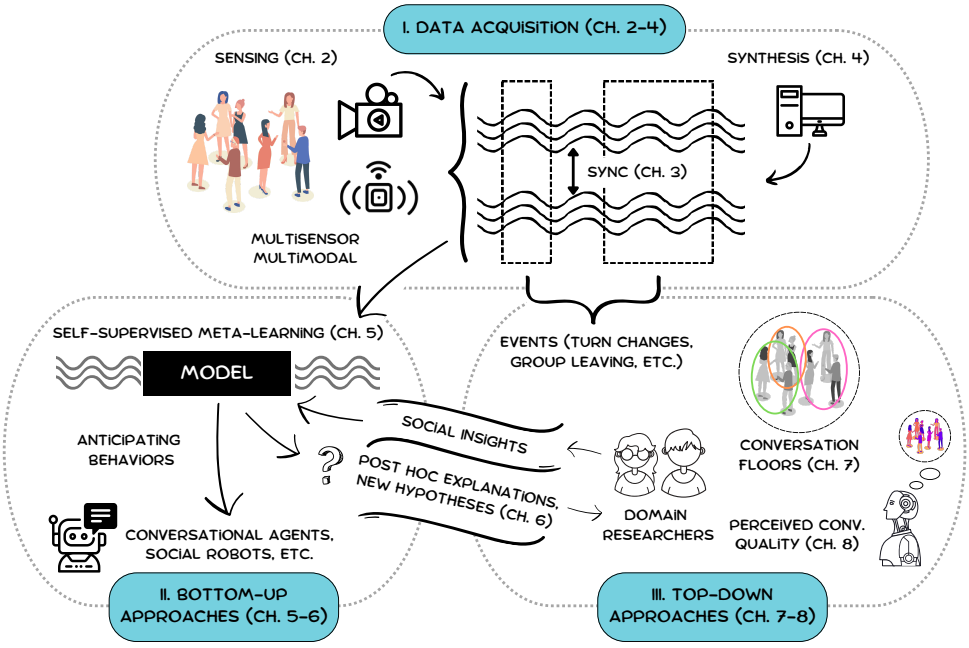


Figure 1.2: Overview of the research contributions presented in this Thesis and how they relate to each other towards advancing Artificial Social Intelligence in the wild.

in widely used behavior datasets perform synchronization as a post-processing step. This fails to provide latency guarantees suitable for studying social phenomena such as mimicry and synchrony which occur at the time scale of tens of milliseconds. Meanwhile, wireless solutions from the broadcasting industry are prohibitively expensive for typical academic research budgets. From a survey of latency measures in social literature, we identify 40 ms as a suitable tolerable latency for human behavior research. Through technical trade-offs, the described approach attains an empirical crossmodal latency of 13 ms at worst, at 1/8-th of the cost of the synchronization solutions used in commercial broadcasting.

The cost and effort involved in collecting even a single dataset in the wild are formidable. So, even with a reproducible template, it is challenging to acquire representative datasets across interaction settings and cultures. Consequently, this Thesis explores whether synthetic data can replace real-world training data for in the wild tasks. The promise of a synthetic data pipeline lies in the control it affords. Real-world data may contain biases along sensitive variables arising from inadequate representation of ethnicities, cultures, appearances, and interaction contexts. These biases are subsequently reflected in models trained on such data. Additionally, synthetic identities also alleviate several privacy and ethical concerns that are relevant when recording real people. Can we synthesize training data to achieve downstream performance comparable with training on real-world data?

Chapter 4 investigates the broader question of whether photorealism is excessive for synthetic training data for face-related computer vision tasks, a domain where synthetic data has already demonstrated promising results. Specifically, we boost the realism of our synthetic faces by introducing dynamic skin wrinkles in response to facial expressions and observe performance improvements in downstream tasks of landmark localization and surface-normal estimation. The key contribution is an approach that produces realistic wrinkles across a large and diverse population of digital humans. We do this by aggregating wrinkling effects directly from high-quality expression scans of people. By leveraging a measure of tension in the face mesh, the proposed method scales with an increasing number of identities and expressions without any additional manual effort and produces realistic wrinkles for expressions not represented in the source scans.

While the synthesis of realistic multimodal social behavior constitutes an overarching motivation, focusing on the more widely studied domain of faces enables the evaluation of using synthetic training data on well-established tasks and benchmarks. Consequently, the chapter represents an important stepping stone toward generating general social behavior across modalities.

1

1.4.2 Modeling: Forecasting & Explaining Social Behavior

Motivated by its success across several domains, a recent trend has been the application of deep learning techniques to various tasks involving nonverbal social human behavior data. However, a crucial consideration here is data efficiency. The scarcity of in-the-wild data is compounded by the fact that the phenomena or events of interest occur infrequently over the duration of the interaction. Examples of such events of interest include speaker turn transitions [72, 73], mimicry episodes [74], disengagement or interaction termination [75], or high-order social actions such as *stepping, laughing, drinking, etc.* [76, 77]. This precludes the training of large networks for modeling such phenomena. A natural question arises: How can we apply deep learning methods in the small-data regime that is social human behavior research? Furthermore, nonverbal behavior is a function of several individual factors such as age, cultural background, and personality variables [78, Ch. 1; 79, p. 237]. How can models adapt their predictions to the idiosyncrasies of individuals and groups? Training a separate model per individual or group of interacting partners would further confound the data scarcity issue.

Chapter 5 follows a bottom-up approach and formalizes the self-supervised task of *Social Cue Forecasting*, with specific task requirements motivated from social science literature. The ability to anticipate behaviors of interacting partners is a critical outcome of the theory of mind, and is consequently a crucial ability towards developing artificial social intelligence. Computationally, the idea is to learn neural representations of general social behavior by leveraging the larger amount of event-agnostic low-level nonverbal cues. Specifically, this is done by forecasting future low-level cues from the same preceding cues over the entire available interaction data. Moreover, taking a meta-learning and stochastic view of group dynamics, we propose the *Social Process* (SP) family of models. At training, SP methods condition their predictions for a sequence on a set of context interaction sequences for a given conversing group, thereby learning to adapt to the dynamics in the context set. In this way, the models can generalize to unseen groups at test by conditioning on a correspondingly unseen context set, avoiding the need to train group-specific models.

While early applications of ASI primarily involved theory- and data-driven *simulations*, much has changed in the landscape of AI research. How can contemporary domain experts develop new social theories given a model that predicts the low-level dynamics of real-world social behavior? Specifically, the next question this Thesis considers is how models forecasting low-level nonverbal cues can be leveraged in forming data-driven hypotheses about causal relationships between high-order social behaviors.

Chapter 6 proposes a post hoc saliency-based explanation framework for counterfactual reasoning in probabilistic multivariate forecasting. The chapter begins by revisiting what constitutes a causal explanation and establishes a conceptual link between counterfactual

reasoning and saliency-based explanation methods. To address the lack of a principled notion of saliency in existing explainable AI methods, we leverage a unifying expression of bottom-up saliency grounded in preattentive human visual cognition and extend it to forecasting settings. The chapter concludes with a case study of how the proposed framework may be used in forming hypotheses surrounding group-leaving behavior using real-world data and forecasting models. In doing so, the broader goal is to bridge bottom-up modeling approaches with the domain insights obtained from top-down approaches.

1.4.3 Perception: Analyzing & Quantifying Social Phenomena

The final part of this Thesis deals with advancing social theories and developing novel measures for quantifying social phenomena within in-the-wild interactions. How can we improve and evaluate existing theories of social interactions in real-world settings? What constructs or instruments are required for quantifying social phenomena in these settings? The research approach here is top-down, following methods more native to traditional social psychology and affective computing.

Chapter 7 deals with identifying conversations in free-standing interactions. Specifically, we unify spatial and temporal notions of a conversation to establish evidence for the presence of multiple simultaneous conversations within a single spatial free-standing conversing group (FCG, operationalized through the framework of F-formations [79]). The chapter begins with Hung's [80] observation that the prior state-of-the-art interpretation of an F-formation assumed only one conversation to exist within it. This assumption did not match our personal observations and experiences. To establish supporting evidence, we visit early conversation analysis literature for the notion of a *conversation floor*, which incorporates the temporal factors in the development of conversations. Using simultaneous speaking turns as a key feature, the chapter establishes empirical evidence for the existence of multiple floors within a single F-formation, and provides post hoc analysis for investigating the effect of group size on speaking turn durations of simultaneous speakers.

Chapter 8 proposes a perceived measure of the quality of spontaneous conversations. Prior research operationalized the quality of conversations in narrow terms, associating greater quality to less small talk. Other works taking a perspective of interaction experience have indirectly studied quality through one of the several overlapping constructs in isolation, such as rapport or engagement. Instead, we propose a holistic conceptualization of conversation quality building upon collaborative attributes of cooperative conversation floors. Specifically, we take a multilevel perspective of conversations: we propose and validate individual- and group-level instruments for capturing external raters' gestalt impressions of participant experiences from thin slices of nonverbal behavior.

1

References

[1] D. Umberson and J. K. Montez. Social Relationships and Health: A Flashpoint for Health Policy. *Journal of health and social behavior*, 51:S54–S66, 2010. doi: 10.1177/0022146510383501.

[2] R. Byrne and A. Whiten. *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford: Clarendon Press, Oxford, UK, 1988.

[3] A. Whiten and R. W. Byrne. *Machiavellian intelligence II: Extensions and evaluations*. Cambridge University Press, 1997.

[4] R. A. Brooks, C. Breazeal, M. Marjanović, B. Scassellati, and M. M. Williamson. The Cog project: Building a humanoid robot. In *Computation for Metaphors, Analogy, and Agents*, pages 52–87. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

[5] R. A. Brooks and L. A. Stein. Building brains for bodies. *Autonomous Robots*, 1:7–25, 1994.

[6] K. Dautenhahn. Getting to know each other—Artificial social intelligence for autonomous robots. *Robotics and Autonomous Systems*, 16(2):333–356, 1995. doi: 10.1016/0921-8890(95)00054-2.

[7] J. Lindblom and T. Ziemke. Social situatedness of natural and artificial intelligence: Vygotsky and beyond. *Adaptive Behavior*, 11(2):79–96, 2003.

[8] W. S. Bainbridge, E. E. Brent, K. M. Carley, et al. Artificial Social Intelligence. *Annual Review of Sociology*, 20:407–436, 1994.

[9] K. M. Carley. Artificial intelligence within sociology. *Sociological Methods & Research*, 25:3–30, 1996.

[10] R. A. Calvo, S. D'Mello, J. M. Gratch, and A. Kappas. *The Oxford handbook of affective computing*. Oxford Library of Psychology, 2015.

[11] E. Aronson, T. D. Wilson, and R. M. Akert. *Social Psychology*. Pearson, Boston, ninth edition edition, 2016.

[12] H. Hung, E. Gedik, and L. Cabrera Quiros. Chapter 11 - complex conversational scene analysis using wearable sensors. In *Multimodal Behavior Analysis in the Wild*, Computer Vision and Pattern Recognition, pages 225–245. Academic Press, 2019. doi: 10.1016/B978-0-12-814601-9.00019-5.

[13] E. L. Thorndike. Intelligence and its uses. *Harper's Magazine*, 140:227–235, 1920.

[14] F. Lievens and D. Chan. Practical Intelligence, Emotional Intelligence, and Social Intelligence. In *Handbook of Employee Selection*, pages 339–359. Routledge, New York, NY : Routledge, 2010. doi: 10.4324/9781315690193-15.

[15] F.-Y. Wang, P. Ye, and J. Li. Social Intelligence: The Way We Interact, The Way We Go. *IEEE Transactions on Computational Social Systems*, 6(6):1139–1146, Dec. 2019. doi: 10.1109/TCSS.2019.2954920.

[16] D. Silvera, M. Martinussen, and T. Dahl. The Tromsø Social Intelligence Scale, a self-report measure of social intelligence. *Scandinavian journal of psychology*, 42:313–9, Oct. 2001. doi:

10.1111/1467-9450.00242.

[17] M. L. Barnes and R. J. Sternberg. Social intelligence and decoding of nonverbal cues. *Intelligence*, 13(3):263–287, 1989.

[18] M. E. Ford and M. S. Tisak. A further search for social intelligence. *Journal of Educational Psychology*, 75(2):196–206, 1983.

[19] D. P. Keating. A search for social intelligence. *Journal of Educational psychology*, 70(2):218, 1978.

[20] R. Strang. Measures of social intelligence. *American Journal of Sociology*, 36(2):263–269, 1930.

[21] D. L. Mackaye. The interrelation of emotion and intelligence. *American Journal of Sociology*, 34 (3):451–464, 1928.

[22] C.-M. T. Wong, J. D. Day, S. E. Maxwell, and N. M. Meara. A multitrait-multimethod study of academic and social intelligence in college students. *Journal of educational psychology*, 87(1): 117, 1995.

[23] J. Williams, S. M. Fiore, and F. Jentsch. Supporting Artificial Social Intelligence With Theory of Mind. *Frontiers in Artificial Intelligence*, 5:750763, Feb. 2022. doi: 10.3389/frai.2022.750763.

[24] F. Cuzzolin, A. Morelli, B. Cîrstea, and B. J. Sahakian. Knowing me, knowing you: Theory of mind in AI. *Psychological Medicine*, 50(7):1057–1061, May 2020. doi: 10.1017/S0033291720000835.

[25] J. Perner and B. Lang. Development of theory of mind and executive control. *Trends in Cognitive Sciences*, 3(9):337–344, Sept. 1999. doi: 10.1016/S1364-6613(99)01362-5.

[26] S. Baron-Cohen and C. Gillberg. Mind blindness: An essay on autism and theory of mind. *Developmental Medicine and Child Neurology*, 37(12):1124–1124, 1995.

[27] H. M. Wellman. Developing a theory of mind. 2011.

[28] A. Senju. Spontaneous theory of mind and its absence in autism spectrum disorders. *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry*, 18(2):108–113, Apr. 2012. doi: 10.1177/1073858410397208.

[29] M. R. A. Chance and A. P. Mead. Social behavior and primate evolution. *Symposia of the Society for Experimental Biology*, 8:395–439, 1953.

[30] A. Jolly. Lemur Social Behavior and Primate Intelligence: The step from prosimian to monkey intelligence probably took place in a social context. *Science (New York, N.Y.)*, 153(3735):501–506, 1966.

[31] N. K. Humphrey. The social function of intellect. *Growing points in ethology*, 37(1):303–317, 1976.

[32] N. J. Emery, N. S. Clayton, and C. D. Frith. Introduction. Social intelligence: From brain to culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):485–488, Apr. 2007. doi: 10.1098/rstb.2006.2022.

[33] A. Kappas, R. Stower, and E. J. Vanman. *Communicating with Robots: What We Do Wrong and What We Do Right in Artificial Social Intelligence, and What We Need to Do Better*, pages 233–254.

Springer International Publishing, Cham, 2020. doi: 10.1007/978-3-030-34964-6_8.

[34] K. Dautenhahn. A Paradigm Shift in Artificial Intelligence: Why Social Intelligence Matters in the Design and Development of Robots with Human-Like Intelligence. In *50 Years of Artificial Intelligence: Essays Dedicated to the 50th Anniversary of Artificial Intelligence*, Lecture Notes in Computer Science, pages 288–302. Springer, Berlin, Heidelberg, 2007. doi: 10.1007/978-3-540-77296-5_26.

[35] T. J. Wiltshire, E. J. C. Lobato, J. Velez, F. G. Jentsch, and S. M. Fiore. An interdisciplinary taxonomy of social cues and signals in the service of engineering robotic social intelligence. In *SPIE Defense + Security*, page 90840F, Baltimore, Maryland, USA, June 2014. doi: 10.1117/12.2049933.

[36] N. Ambady, F. J. Bernieri, and J. A. Richeson. Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In *Advances in experimental social psychology*, volume 32, pages 201–271. Elsevier, 2000.

[37] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[38] I. Oguntola, D. Hughes, and K. Sycara. Deep Interpretable Models of Theory of Mind, July 2021.

[39] J. Hoey, T. Schröder, J. Morgan, et al. Artificial Intelligence and Social Simulation: Studying Group Dynamics on a Massive Scale. *Small Group Research*, 49(6):647–683, Dec. 2018. doi: 10.1177/1046496418802362.

[40] J. Mökander and R. Schroeder. AI and social theory. *AI & SOCIETY*, 37(4):1337–1351, Dec. 2022. doi: 10.1007/s00146-021-01222-z.

[41] J. Radford and K. Joseph. Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science. *Frontiers in Big Data*, 3, 2020.

[42] K. Dautenhahn, B. Ogden, and T. Quick. A framework for the study of socially embedded and interaction-aware robotic agents. *Cognitive Systems Research*, 3(3):397–428, 2002.

[43] P. Sengers. Socially situated AI: What it means and why it matters. In *Proceedings of the 1996 AAAI Symposium, Entertainment and AI/A-Life. Technical Report WS-96-03*, pages 69–75, 1996.

[44] H. Gardner. *The mind's new science: A history of the cognitive revolution.* Basic books, 1987.

[45] A. Turing. Computing machinery and intelligence. *Mind*, 49:433–460, 1950.

[46] R. M. French. Subcognition and the limits of the turing test. *Mind*, 99(393):53–65, 1990.

[47] B. Edmonds. The Social Embedding of Intelligence: Towards Producing a Machine that Could Pass the Turing Test. In *Parsing the Turing Test*, pages 211–235. Springer Netherlands, Dordrecht, 2009. doi: 10.1007/978-1-4020-6710-5_14.

[48] R. Krishna, D. Lee, L. Fei-Fei, and M. S. Bernstein. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences*, 119(39):e2115730119, Sept. 2022. doi: 10.1073/pnas.2115730119.

[49] S. Bolotta and G. Dumas. Social Neuro AI: Social Interaction as the "Dark Matter" of AI. *Frontiers in Computer Science*, 4, 2022.

[50] A. Turing. Intelligent machinery. 1948.

[51] J. J. Walczyk, K. T. Mahoney, D. Doverspike, and D. A. Griffith-Ross. Cognitive lie detection: Response time and consistency of answers as cues to deception. *Journal of Business and Psychology*, 24:33–49, 2009.

[52] A. M. Treisman and G. Gelde. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[53] J. Saghaei, A. Fallahzadeh, and T. Saghaei. Vapor treatment as a new method for photocurrent enhancement of UV photodetectors based on ZnO nanorods. *Sensors and Actuators A: Physical*, 247:150–155, 2016.

[54] J. R. Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6):643–662, 1935.

[55] I. Biederman, A. L. Glass, and E. W. Stacy. Searching for objects in real-world scenes. *Journal of Experimental Psychology*, 97(1):22–27, 1973.

[56] G. L. Stewart, K. A. Manges, and M. M. Ward. Empowering sustained patient safety: the benefits of combining top-down and bottom-up approaches. *Journal of Nursing Care Quality*, 30(3): 240–246, 2015.

[57] S. Cohen. *The nature of moral reasoning: The framework and activities of ethical deliberation, argument and decision making*. Oxford University Press, 2004.

[58] C. P. Lynam, M. Llope, C. Möllmann, et al. Interaction between top-down and bottom-up control in marine food webs. *Proceedings of the National Academy of Sciences*, 114(8):1952–1957, 2017.

[59] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22 (12):1349–1380, 2000.

[60] A. B. Costello and J. Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, 10 (7), 2005.

[61] J. A. Allen, C. Fisher, M. Chetouani, et al. Comparing social science and computer science workflow processes for studying group interactions. *Small Group Research*, 48(5):568–590, 2017. doi: 10.1177/1046496417721747.

[62] C. C. S. Liem, M. Langer, A. Demetriou, et al. Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 197–253. Springer International Publishing, Cham, 2018. doi: 10.1007/978-3-319-98131-4_9.

[63] E. Nilsen, D. Bowler, and J. Linnell. Exploratory and confirmatory research in the open science era. *Journal of Applied Ecology*, 57, Feb. 2020. doi: 10.1111/1365-2664.13571.

[64] K. R. Hammond. Ecological validity: Then and now, 1998.

[65] J. F. Kihlstrom. Ecological validity and "ecological validity". *Perspectives on Psychological Science*, 16(2):466–471, 2021.

[66] E. Aronson and J. M. Carlsmith. Experimentation in social psychology. *The handbook of social psychology*, 2(2):1–79, 1968.

[67] Y. Rogers and P. Marshall. Research in the wild. *Synthesis Lectures on Human-Centered Informatics*, 10(3):i–97, 2017.

[68] T. Stadelmann, M. Amirian, I. Arabaci, et al. Deep learning in the wild. In *Artificial Neural Networks in Pattern Recognition: 8th IAPR TC3 Workshop, ANNPR 2018, Siena, Italy, September 19–21, 2018, Proceedings 8*, pages 17–38. Springer, 2018.

[69] C. Li, H. Liu, L. H. Li, et al. ELEVATER: a benchmark and toolkit for evaluating language-augmented visual models. In *36th Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

[70] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge (EmotiW) challenge and workshop summary. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pages 371–372. Association for Computing Machinery, 2013.

[71] H. Hung, E. Gedik, and L. C. Quiros. Complex conversational scene analysis using wearable sensors. In *Multimodal Behavior Analysis in the Wild*, pages 225–245. Elsevier, 2019.

[72] S. Garrod and M. J. Pickering. The use of content and timing to predict turn transitions. *Frontiers in psychology*, 6:751, 2015.

[73] A. Keitel and M. M. Daum. The use of intonation for turn anticipation in observed conversations without visual signals as source of information. *Frontiers in psychology*, 6:108, 2015.

[74] S. Bilakhia, S. Petridis, and M. Pantic. Audiovisual Detection of Behavioural Mimicry. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 123–128, Geneva, Switzerland, Sept. 2013. IEEE. doi: 10.1109/ACII.2013.27.

[75] D. Bohus and E. Horvitz. Managing Human-Robot Engagement with Forecasts and… um… Hesitations. *Proceedings of the 16th International Conference on Multimodal Interaction*, page 8, 2014.

[76] L. Airale, D. Vaufreydaz, and X. Alameda-Pineda. SocialInteractionGAN: Multi-person Interaction Sequence Generation. Mar. 2021.

[77] N. Sanghvi, R. Yonetani, and K. Kitani. Mgpi: A computational model of multiagent group perception and interaction. *arXiv preprint arXiv:1903.01537*, 2019.

[78] N.-J. Moore, H. Mark III, and W. Don. Stacks. Nonverbal communication: Studies and applications. 2013.

[79] A. Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP

Archive, 1990.

[80] H. Hung. MINGLE: modelling social group dynamics and interaction quality in complex scenes using multi-sensor analysis of non-verbal behaviour. URL https://www.nwo.nl/en/projects/639022606. Vidi, 18-September-2017 to 17-November-2023.

I

DATA ACQUISITION

Sensing & Synthesis

2

Curating Datasets of Social Interactions in the Wild: Data Collection for the Community by the Community

C. Raman*, J. Vargas-Quiros*, S. Tan*, A. Islam, E. Gedik, and H. Hung. ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild. *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, New Orleans, USA, 2022.

*Equal contribution

Abstract

Recording the dynamics of unscripted human interactions in the wild is challenging due to the delicate trade-offs between several factors: participant privacy, ecological validity, data fidelity, and logistical overheads. To address these, following a datasets for the community by the community ethos, we propose the Conference Living Lab (ConfLab): a new concept for multimodal multisensor data collection of in-the-wild free-standing social conversations. For the first instantiation of ConfLab described here, we organized a real-life professional networking event at a major international conference. Involving 48 *conference attendees, the dataset captures a diverse mix of status, acquaintance, and networking motivations. Our capture setup improves upon the data fidelity of prior in-the-wild datasets while retaining privacy sensitivity:* 8 *videos (*$1920 \times 1080, 60$ *fps) from a non-invasive overhead view, and custom wearable sensors with onboard recording of body motion (full* 9*-axis IMU), privacy-preserving low-frequency audio (*1250 *Hz), and Bluetooth-based proximity. Additionally, we developed custom solutions for distributed hardware synchronization at acquisition, and time-efficient continuous annotation of body keypoints and actions at high sampling rates. Our benchmarks showcase some of the open research tasks related to in-the-wild privacy-preserving social data analysis: keypoints detection from overhead camera views, skeleton-based no-audio speaker detection, and F-formation detection.*

2.1 Introduction

A crucial challenge towards developing artificial socially intelligent systems is understanding how *real-life* situational contexts affect social human behavior [1]. Social-science findings indeed show that the dynamics of how we conduct daily interactions vary significantly depending on the social situation [2–4]. Unfortunately, such dynamics are not adequately captured by many data collection setups where role-played or scripted scenarios are typical [5].

In this paper we address the problem of collecting a privacy-sensitive dataset of unscripted social dynamics of real-life relationships where encounters can influence someone's daily life. We argue that doing so requires recording these exchanges in the natural ecology, requiring an approach different from the typical setup of locally-organized studies. Specifically, we focus on free-standing interactions within the setting of an international conference (see Figure 2.1).

Recording an international community in its natural habitat is characterized by several intersecting challenges: an intrinsic trade-off exists between data fidelity, ecological validity, and privacy preservation. For ecological validity, a non-invasive capture setup is essential for mitigating any influence on behavior naturalness [6–8]. The most common solution involves mounting cameras from aerial perspectives such as top-down [9, 10] and elevated-

Figure 2.1: Snapshot of the interaction area from our cameras. We annotated only cameras highlighted with red borders (high scene overlap). For a clearer visual impression of the scene, we omit cameras 1 (few people recorded) and 5 (failed early in the event). Faces blurred to preserve privacy.

side views [11–13]. Now elevated-side views make it easy to capture sensitive personal information such as faces, which leads to several ethical concerns. For instance, capturing faces has been related to harmful downstream surveillance applications [14]. Besides, state-of-the-art (SOTA) body-keypoint estimation techniques perform poorly on aerial perspectives [9, 15], making the extraction of automatic pose annotations challenging (Figure 2.3). To avoid such issues, some researchers have turned to more privacy-preserving wearable sensors shown to benefit many behavior analysis tasks [8, 16, 17].

In all, the closest related datasets (see Table 2.1) suffer from several technical limitations precluding the analysis and modeling of fine-grained social behavior: (i) lack of articulated pose annotations; (ii) a limited number of people in the scene, preventing complex interactions such as group splitting/merging behaviors, and (iii) an inadequate data sampling-rate and synchronization-latency to study time-sensitive social phenomena [18, Sec. 3.3].

To address all these limitations, we propose the Conference Living Lab (ConfLab): a new concept for multimodal multisensor data collection of ecologically-valid social settings. From the first instantiation of ConfLab, we provide a high-fidelity dataset of 48 participants at a professional networking event.

Methodological Contributions: We describe a data collection design that captures a diverse mix of real levels of seniority, acquaintance, affiliation, and motivation to network (see Figure 2.2). This was achieved by organizing ConfLab as part of a major international scientific conference. ConfLab had these goals: (i) a data collection effort follwing a *by the community for the community* ethos: the more volunteers, the more data, (ii) volunteers who potentially use the data can experience first-hand potential privacy and ethical considerations related to sharing their own data, (iii) in light of recent data sourcing issues [14, 19], we incorporated privacy and invasiveness considerations directly into the decision-making process regarding sensor type, positioning, and sample-rates.

Technical Contributions: **(i) aerial-view articulated pose**: our annotations of 17 full-body keypoints enable improvements in (a) pose estimation and tracking, (b) pose-based

Table 2.1: Comparison of ConfLab with prior datasets of free-standing conversation groups in in-the-wild social interaction settings. Conflab is the first and only social interaction dataset that offers skeletal keypoints and speaking status at high annotation resolution, as well as hardware synchronized camera and multimodal wearable signals at high resolution.

Dataset	People/ Scene	Video	Manual Annotations	Wearable Signals	Synchronization
Cocktail [13]†	7	512 × 384	F-formations (20 and 30 min, 1/5 Hz)	None	Unknown
CoffeeBreak [12]	14	1440 × 1080	F-formations (130 frames in two sequences)	None	None
IDIAP [10]	> 50	180 min; 654 × 439 20 fps	F-formations (82 independent frames)	None	None
SALSA [11]†	18	60 min; 1024 × 768 15 fps	Bounding boxes (30 min) Head & body ori. (30 min) F-formations (60 min) (all 1/3 Hz)	Audio MFCCs (30 Hz) Acceleration (20 Hz) IR proximity (1 Hz)	Post-hoc infra-red event-based (no-drift assumption)
MnM [9]†	32	30 min; 1920 × 1080 30 fps	Bounding boxes (30 min, 1 Hz ‡) F-formations (10 min, 1 Hz) Actions (45 min, 1 Hz‡)	Accelerometer (20 Hz) Radio proximity (1 Hz)	Intra-wearable sync via gossiping protocol; Inter-modal sync using manual inspection @1 Hz
ConfLab	48	~ 45 min; 1920 × 1080 **60 fps**	**17 keypoints (16 min, 60 Hz)** F-formations (16 min, 1 Hz) **Speaking status (16 min, 60 Hz)**	**Low-freq. audio (1250 Hz)** **BT proximity (5 Hz)** **9-axis IMU (56 Hz)**	**Wireless hardware sync at acquisition, max latency of ~ 13 ms** [18]

† Includes self-assessed personality ratings ‡ Upsampled to 20 Hz using Vatic [20]
BT: Bluetooth IMU: Inertial Measurement Unit

recognition of social actions (under-explored in the top-down perspective), (c) pose-based F-formation estimation (has not been possible from prior work [10, 21–23]), and (d) the direct study of interaction dynamics using full body poses (previously limited to lab settings [24]). **(ii) subtle body dynamics**: we are the first to use a full 9-axis Inertial Measurement Unit (IMU) enabling a richer representation of behaviour at higher sample rates; previous rates were found to be insufficient for downstream tasks [17]. **(iii) enabling finer temporal-scale research questions**: a sub-second crossmodal latency of ~ 13 ms along with higher sampling rate of features (60 fps video, 56 Hz IMU) opens the gateway for the in-the-wild study of nuanced time-sensitive social behaviors like mimicry and synchrony.

2.2 Related Work

Early datasets of in-the-wild social events either spanned only a few minutes (e.g. Coffee Break [12]), or were recorded at such a large distance from the participants that performing robust, automated person detection or tracking with SOTA approaches was non-trivial (e.g. Idiap Poster Data [10]). More recently, two different strategies have emerged to circumvent

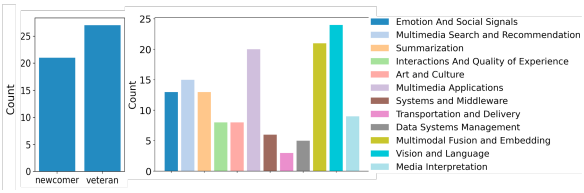


Figure 2.2: Frequency of newcomer/veteran participants (left) and reported research interests (right).

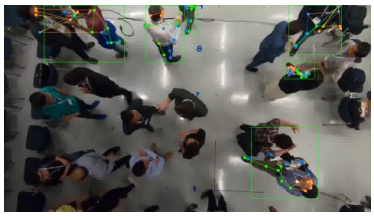


Figure 2.3: Keypoint detection using pre-trained RSN [28]. Additional SOTA results are in Appendix 2.F.1

such issues.

One approach involves fully instrumented labs with a high resolution multi-camera setup for video and audio data. Here automatic detectors [24–26] could be applied to obtain poses. This circumvents the cost- and labor-intensive process of manually labeling head poses, at the cost of less portable sensing setups. Notable examples of such in-the-lab studies include seated scenarios, such as the AMI meeting corpus [27], and more recently standing scenarios like the Panoptic Dataset [24]. Both enable the learning of multimodal behavioral dynamics. However, the dynamics of seated, scripted, or role-playing scenarios are different from that of an unconstrained social setting such as ours. In contrast, ConfLab moves out of the lab with a more modular and portable multimodal, multisensor solution that scales easily in the wild.

Another approach exploited wearable sensor data to allow for multimodal processing—sensors included 3 or 6 DOF inertial measurement units (IMU); infrared, bluetooth, or radio sensors to measure proximity; or microphones for speech behavior [9, 11]. While proximity has been used as a proxy of face-to-face interaction [11, 29–32], recent findings highlight significant problems with such an assumption [33]. Such errors can have a significant impact on the machine-perceived experience of an individual, precluding the development of personalized technology. Chalcedony badges used by [9] show more promising results with a radio-based proximity sensor and accelerometer [34], but such data remains insufficient for more downstream tasks due to the relatively low sample (20Hz) and annotation (1Hz) frequency [17]. In light of these challenges in wearable sensing, ConfLab features custom-developed Midge sensors that enable more flexible and fine-grained on-device recording. At the same time, ConfLab enables researchers in the wearable and ubiquitous computing communities to investigate the benefit of exploiting wearable and multimodal data.

Furthermore, while both SALSA [11] and MatchNMingle [9] capture a multimodal dataset of a large group of individuals involved in mingling behavior, the inter-modal synchronization is only guaranteed at 1/3 Hz and 1 Hz, respectively. Prior works coped

with lower tolerances by computing summary statistics over input windows [17, 35, 36]. While 1 Hz is able to capture some conversation dynamics [37], it is insufficient to study fine-grained social phenomena such as back-channeling or mimicry that involve far lower latencies [18, Sec. 3.3]. ConfLab provides data streams with higher sampling rates, synchronized at acquisition with our method shown to yield a 13 ms latency at worst [18] (see Section 2.3). Table 2.1 summarizes the differences between ConfLab and other related datasets.

2.3 Data Acquisition

In this section we describe the considerations, design, and supporting community engagement activities for the first instantiation of ConfLab at ACM Multimedia 2019 (MM'19), to serve as a template and case study for other similar efforts.

Ecological Validity and Recruitment An often-overlooked but crucial aspect of in-the-wild data collection is the design and ecological validity of the interaction setting [6–8]. To capture natural interactions in a professional setting and encourage mixed levels of status, acquaintance, and motivations to network, we co-designed a networking event with the MM'19 organizers called *Meet the Chairs!* Our event website (https://conflab.ewi.tudelft.nl/) served to inform participants about the goals of a community created dataset, and transparently describe the data collection process (Figure 2.4). During the conference, participants were recruited via word-of-mouth marketing, social media, conference announcements, and the event website. As an additional incentive beyond interacting with the Chairs and participating in a community-driven data endeavor, we provided attendees with post-hoc insights into their networking behavior from the collected wearable-sensors data. See Supplementary material for a sample participant report.

Privacy and Ethics The collection and sharing of ConfLab is GDPR compliant. The dataset design and process was approved by both, the Human Research Ethics Committee (HREC) at our institution (TUDelft) and the conference location's national authorities (France). All participants gave consent for the recording and sharing of their data at regis-

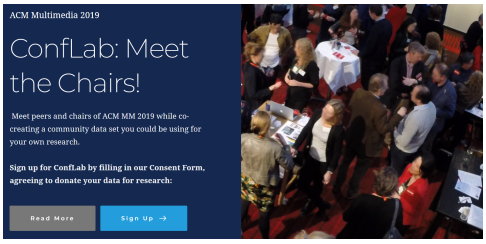


Figure 2.4: Screenshots from the *ConfLab: Meet the Chairs!* event website

Figure 2.5: The Midge

tration. (See the Datasheet in the Appendix for the consent form.) Given the involvement of private human data, ConfLab is only available for academic research purposes under an End User License Agreement. Such an *as open as possible and as closed as necessary* ethos for open science acknowledges the limitation that personal data places on open sharing [38, 39].

Data Capture Setup Our goal while designing the capture setup was to find the best trade-off between maximizing data fidelity and interfering with the naturalness of the interaction (ecological validity) or violating participant privacy (ethical considerations). Through discussions with the HREC and General Chairs of MM'19 we decided to mitigate the capture of faces, which constitute one of the most sensitive personally-identifiable features. Avoiding the inclusion of faces serves two purposes. First, it safeguards against misuse in downstream tasks with potential negative societal impacts such as harmful surveillance. Such issues have led to the retraction of some person re-identification datasets [14]. Second, it protects the participants who are part of a real research community; since the dataset does not involve role-playing or scripted conversations, the dataset contains their actual behavior. Consequently, we chose an aerial perspective for the video modality (see Figure 2.6). The 10 m × 5 m interaction area was recorded by 14 GoPro Hero 7 Black cameras (60fps, 1080p, Linear, NTSC) [40]. 10 of these were placed directly overhead at a height of ~ 3.5 m at 1 m intervals, with 4 cameras at the corners providing an elevated-side-view perspective. (The HREC has suggested not sharing the elevated-side-view videos due to the presence of faces.) For capturing multimodal data streams, we designed a custom wearable multi-sensor pack called the Midge[1] (see Figure 2.5 for a design render), based on the open-source Rhythm Badge designed for office environments [41]. We improved upon the Rhythm Badge to achieve more fine-grained and flexible data capture (see Appendix 2.D). We designed the Midge in a conference badge form-factor for seamless integration. Unlike smartphones, wearable badges allow for a simple *grab-and-go* setup and do not suffer from sensor/firmware differences across models. Popular human behavior datasets are synchronized by maximizing similarity scores around manually identified common events, such as infrared camera detections [11], or speech plosives [42]. While recordings in lab settings can allow for fully wired recording setups, recording in-the-wild requires a distributed wireless solution. We developed a solution to synchronize the cameras and wearable sensors directly at acquisition while significantly lowering the cost of the recording setup [18], making it easier for others to replicate our capture setup. See Appendix 2.D for synchronization and calibration details, and Appendix 2.B for images of the setup.

[1]Documentation and schematics: https://github.com/TUDelft-SPC-Lab/spcl_midge_hardware

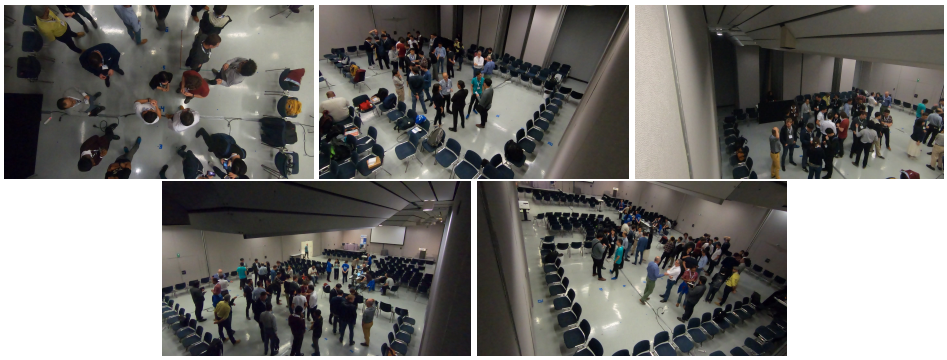


Figure 2.6: Comparing the top-down (top-left, camera 4) and elevated-side camera views (rest). Note how the top-down view is better at mitigating the capture of faces and suffers from fewer occlusions. This allows for a clearer capture of gestures and lower extremities for the most number of people while also preserving privacy.

Data Association and Participant Protocol One consideration for multimodal data recording is the data association problem—how can pixels corresponding to an individual be linked to their other data streams? To this end, we designed a participant registration protocol. Arriving participants were greeted and fitted with a Midge. The ID of the Midge acted as the participant's identifier. One team member took a picture of the participant while ensuring both the face of the participant and the ID on the Midge were visible. In practice, it is preferable to avoid this step by using a fully automated multimodal association approach. However this remains an open research challenge [43, 44]. During the event, participants mingled freely—they were allowed to carry bags or use mobile phones. Conference volunteers helped to fetch drinks for participants. Participants could leave before the end of the one hour session.

Replicating Data Collection Setup and Community Engagement After the event, we gave a tutorial at MM'19 [45] to demonstrate how our collection setup could be replicated, and to invite conference attendees and event participants to reflect on the broader considerations surrounding privacy-preserving data capture, sharing, and future directions such initiatives could take.

2.4 Data Annotation

Continuous Keypoints Annotation Existing datasets of in-the-wild social interactions have mainly focused on localizing subjects via bounding boxes [9, 11]. However, richer information about the social dynamics such as gestures and changes in orientation cannot be retrieved from bounding boxes alone, and necessitates the labeling of multiple skeletal keypoints. The typical approach to keypoint annotation involves using tools such as Vatic [20] or CVAT [46] to manually label every N frames followed by interpolating over the rest

of the frames. This one-frame-at-a-time annotation procedure makes obtaining keypoint annotations a labor- and cost-intensive process. Moreover, interpolation fails to capture the finer temporal dynamics of the underlying behavior, and reduces the benefits of higher-framerate video capture. Limited by existing tools, no related dataset of in-the-wild human behavior has included time-continuous pose or speaking status annotations.

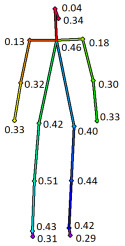
In contrast, to overcome these issues we collected fine-grained time-continuous annotations of keypoints via a web-based interface implemented as part of the Covfee framework [47]. Here, annotators follow individual joints using their mouse or trackpad while playing the video in their web browser. The playback speed of the video is automatically adjusted using an optical-flow-based technique to enable annotators to follow keypoints continuously without pausing the video. This design enables easy keypoint labeling in *every* frame of the video (60 Hz). We also incorporated a binary *occlusion* flag for every body keypoint. Annotators simultaneously controlled this flag to indicate when a body joint was not directly visible. Note that the flag is only an additional confidence indicator; we asked the annotators to label the occluded keypoint using their best estimate if it was deemed to be within the frame. Our pilot study on the efficacy of Covfee compared to non-continuous annotation via CVAT [46] is presented in [47]. For the pilot annotators, the continuous annotation methodology resulted in a 3× speedup with statistically indifferent error rates.

We chose the top-down camera views for annotation since they suffer from fewer occlusions than the elevated-side views, enabling improved capture of gestures and lower extremities for more number of people (see Figure 2.6). Given the overlap in the camera views, we annotated keypoints in five of the ten overhead cameras (see Figure 2.1). Note that the same subject could be annotated in multiple cameras due to the overlap in even the five annotated cameras. Videos were split into two-minute segments to ease the annotation procedure. Each segment was annotated by one annotator by tracking the joints of all the people in the scene.

Continuous Speaking Status Annotations Speaking status is a key non-verbal cue for many social interaction analysis tasks [48]. We annotated the binary speaking status of every subject due to its importance as a key feature of social interaction [16, 49–52] and to contribute the existing community who are working on this task [17, 53, 54]. Action annotations have traditionally been carried out using frame-wise techniques [9], where annotators find the start and end frame of the action of interest using a graphical interface. Given the speed enhancement of continuous annotation, we also annotated speaking status via a continuous technique. We implemented a binary annotation interface as part of Covfee [47]. We asked annotators to press a key when they perceived speaking starting or ending. In a pilot study with two annotators, we measured a frame-level agreement (Fleiss' κ) of 0.552, comparable to previous work [35]. Similar to [9], the annotations were

(a) Keypoint annotation interface in covfee [55]

(b) Gallery of identities



(c) Occlusion

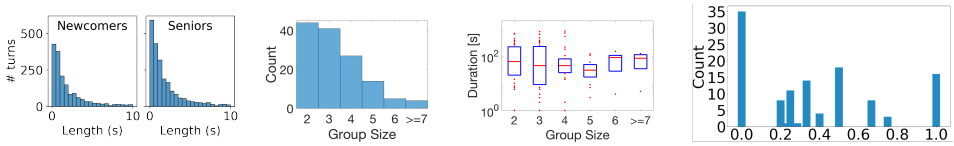
Figure 2.7: Illustration of the body keypoints annotation procedure: (a): our custom time continuous annotation interface; (b): the gallery of person identities used by annotators to identify people in the scene (faces blurred); and (c): the skeleton template with the fraction of occluded frames.

made by watching the video. We provided the annotators with all overhead views to best capture visual behavior.

F-formation Annotations Identifying who is likely to have social influence on whom is another important feature for analyzing social behavior. This is operationalised via the theory of F-formations, which are groups of people arranging themselves to converse or socially interact. Similar to prior datasets [9, 11, 13], F-formations group membership were annotated using an approximation of Kendon's definition [56]. F-formation stands for Facing formation, which is a socio-spatial arrangement where people have direct, easy and equal access while excluding the space from others in the surroundings. The arrangement commonly maintains a convex space in the middle of all the participants (determined by the location and orientation of their lower body), although other spatial arrangements (e.g., side-by-side, L-shaped) are possible, especially for smaller-sized groups of people. Annotations were labeled by one annotator at 1 Hz, following this definition. Since this is a largely objective and common framework for defining F-formations, we deemed it sufficient to obtain one set of annotations. Further, since F-formations may span camera views, we always used the camera that captured each F-formation in its entirety for annotation.

2.5 Dataset Statistics

Individual-Level Statistics Figure 2.7c shows the average occlusion values we obtained from annotators for each of the 17 keypoints. In Figure 2.8a we show the distribution of turn lengths in our speaking status annotations, for both newcomers and veterans, as per their self-reported newcomer status to the conference. We defined a turn to be a contiguous segment of positively-labeled speaking status, which resulted in a total of 4096 turns annotated.



(a) speaking turn lengths **(b)** group size **(c)** group duration **(d)** fraction of newcomers in groups

Figure 2.8: Data distributions for speaking status and conversation groups

Group-Level Statistics We found 119 distinct F-formations of size greater than or equal to two, and 38 instances of singletons. Of these, there are 14 F-formations and 2 singletons that include member(s) using the mobile phone. The distributions for group size and duration per group size are shown in Figure 2.8b and Figure 2.8c, respectively. Mean group duration doesn't seem to be influenced by group size although higher variations are seen at smaller group sizes. The fraction of community newcomers (first-time attending the conference) in groups is summarized in histogram in Figure 2.8d. The figure demonstrates two peaks on both sides of the spectrum (i.e., no newcomers vs. all newcomers in the same group). This spread over mixed and non-mixed seniority presents opportunities to study how acquaintance and seniority influence conversation dynamics.

2.6 Research Tasks

We report experimental results on three baseline benchmark tasks: person and keypoints detection, speaking status detection, and F-formation detection. The first task is a fundamental building block for automatically analyzing human social behaviors. The other two demonstrate how learned body keypoints can be used in the behavior analysis pipeline. We chose these benchmarking tasks since they have been commonly studied on other in-the-wild behavior datasets. Code for all benchmark tasks is available at: https://github.com/TUDelft-SPC-Lab/conflab. See the *Uses* section of the Datasheet in the Appendix for a discussion of the broader range of tasks ConfLab enables.

2.6.1 Person and Keypoints Detection

This benchmark involves the tasks of person detection (identifying bounding boxes) and pose estimation (localizing skeletal keypoints). Since pre-trained SOTA methods struggle with a privacy-sensitive top-down perspective [15] (also see Figure 2.3 and Appendix 2.F.1 for ConfLab results), we finetune COCO-pretrained models on our dataset. We used Mask-RCNN [57] (Detectron2 framework [58] implementation) with a ResNet-50 backbone for both tasks for benchmarking. Since keypoint annotations were made per camera, we used four of the overhead cameras for training (Cameras 2, 4, 8, 10) and one for testing (Camera 6). Implementation details are available in Appendix 2.E.1.

Table 2.2: Mask-RCNN results for person bounding box detection and keypoint estimation.

Model	Person Detection			Keypoint Estimation		
	AP_{50}	AP	AP_{75}	AP^{OKS}_{50}	AP^{OKS}	AP^{OKS}_{75}
R50-FPN	73.9	38.9	38.4	45.3	13.5	3.3

Figure 2.9: Predictions from the Mask-RCNN model; COCO pretrained (left), and ConfLab finetuned (right).

Evaluation Metrics We evaluated person-detection performance using the standard metrics in the MS-COCO dataset paper [59]. We report average precision (AP) for intersection over union (IoU) thresholds of 0.50 and 0.75, and the mean AP from an IoU range from 0.50 to 0.95 in 0.05 increments. For keypoint detection, we use object keypoint similarity (OKS) [59]. AP^{OKS} is a mean average precision for different OKS thresholds from 0.5 to 0.95.

Results and Analyses Table 2.2 summarizes our person detection and joint estimation results. Our baseline achieves 73.9 AP_{50} in detection and 45.3 AP^{OKS}_{50} in keypoint estimation. Figure 2.9 shows qualitative results from our fine-tuned network.

For further insight we performed several analyses and ablations. In Appendix Table 2.6, we depict the effect of varying the number of training samples on performance. For training, we use the same four cameras and only vary the number of frames for each camera. We evaluate on the same testing images from camera 6. We find that performance saturates at 16% training samples. We next investigated the effect of increasing training data size by adding specific cameras one at a time. We report results in Appendix Table 2.7. There is a 260% performance gain when first doubling the training samples to 69 k with the addition of camera 4, and a 46% gain when adding another 43 k samples from camera 8. Finally, since the lower body regions suffer from higher occlusion, we experiment with different sections of body for further insight and report results in Appendix Table 2.8.

2.6.2 Speaking Status Detection

In data collected from real-life social settings, individual audio recordings can be hard to obtain due to privacy concerns [60]. This has led to the exploration of other modalities to capture some of the motion characteristics of speaking-related gestures [35, 36]. In this task we explore the use of body pose and wearable acceleration data for detecting the speaking status of a person in the scene.

Setup We use the SOTA MS-G3D graph neural network for skeleton action recognition [61], pre-trained on Kinetics Skeleton 400. For the acceleration modality, we evaluated three time series classifiers, each of which we trained from scratch: 1D Resnet [62], InceptionTime [63], and Minirocket [64]. We performed late fusion by averaging the scores from both modalities. Like prior work [17, 36], the task was set up as a binary classification problem.

We divided our pose (skeleton) tracks into 3-second windows with 1.5 s overlap. A window was labeled positive if more than 50% of the continuous speaking status labels within it are positive. This resulted in an imbalanced dataset of 42882 windows with 29.2% positive labels. Poses were pre-processed for training following [61]. Three of the keypoints (head, and feet tips) were discarded due to not being present in Kinetics. We adapted the network by freezing all layers except for the last fully connected layer and training for five extra epochs. Acceleration readings were not pre-processed, other than by interpolating the original variable-sampling-rate signals to a fixed 50 Hz.

Evaluation Evaluation was carried out via 10-fold cross-validation at the subject level, ensuring that no examples from the test subjects were used in training. We used the area under the ROC curve (AUC) as main evaluation metric to account for the imbalance in the labels.

Results The results in Table 2.3 indicate a better performance from the acceleration-based methods. One possible reason for the lower performance of the pose-based methods is the significant domain shift between Kinetics and Conflab, especially in camera viewpoint (frontal vs top-down). The acceleration performance is in line with previous work [17]. Multimodal results were slightly higher than acceleration-only results, despite our naive fusion approach, a possible point to improve in future work [65]. Experiments with the rest of the IMU modalities are presented in Appendix 2.F.2.

2.6.3 F-formation Detection

Setup Like prior work [10, 21–23], we operationalize interaction groups using the framework of F-formations [56]. We provide performance results for F-formation detection using GTCG [23] and GCFF [67] as a baseline. Recent deep learning methods such as DANTE [22] are not directly applicable since they depend on knowing the number of people in the scene, which is variable for ConfLab. We use pre-trained model parameters (reported in the original GTCG and GCFF papers on the Cocktail Party dataset [13]) and tuned a subset of parameters more relevant to ConfLab attributes on camera 6. More details can be found

Table 2.3: ROC AUC and accuracy of skeleton-based, acceleration-based and multimodal speaking status detection (10-fold cross-validation).

Modality	Model	AUC	Acc.
Pose	MS-G3D [66]	0.676	0.677
	InceptionTime [63]	0.798	0.768
Acceleration	Resnet 1D [62]	0.801	0.767
	Minirocket [64]	0.813	0.768
Multimodal	MS-G3D + Minirocket	0.823	0.775

Table 2.4: Average F1 scores for F-formation detection comparing GTCG [23] and GCFF [67] with the effect of different threshold and orientations (standard deviation in parenthesis).

	GTCG		GCFF	
	T=2/3	T=1	T=2/3	T=1
Head	0.51 (0.09)	0.40 (0.12)	0.47 (0.07)	0.31 (0.23)
Shoulder	0.46 (0.11)	0.38 (0.11)	0.56 (0.25)	0.36 (0.16)
Hip	0.45 (0.10)	0.37 (0.12)	0.39 (0.06)	0.25 (0.11)

in Appendix 2.E.2. We derive three different sets of orientation features from (i) head, (ii) shoulder and (iii) hip keypoints.

Evaluation Metrics We use the standard F1 score as evaluation metric for group detection [23, 67]. A group is correctly estimated (true positive) if at least $\lceil T*|G| \rceil$ of the members of group G are correctly identified, and no more than $1 - \lceil T*|G| \rceil$ is incorrectly identified, where T is the tolerance threshold. We report results for $T = \frac{2}{3}$ and $T = 1$ (more strict threshold) in Table 2.4.

Results We show that different results are obtained using different sources of orientations. Different occlusion levels in keypoints due to camera viewpoint may have affected performance. Another factor influencing model performance is that F-formations (which are driven by lower-body orientations [56]) may have multiple conversations floors [50]. Floors are indicated by coordinated speaker turn taking patterns and influence coordinated head orientations of the group.

2.7 Conclusion and Discussion

ConfLab contributes a new concept for real-life data collection in the wild and captures a high-fidelity dataset of mixed levels of acquaintance, seniority, and personal motivations.

ConfLab: the Dataset We improved upon prior work by providing higher-resolution, fidelity, and synchronization across sensor networks. We also carefully designed our social interaction setup to enable a diverse mix of seniority, acquaintanceship, and motivations for mingling. The result is a rich set of 17 body-keypoint annotations of 48 people at 60 Hz from overhead cameras for developing more robust estimation of keypoints, speaking status and F-formations for further analyses of more complex socio-relational phenomena. Our benchmark results for these tasks highlight how the improved fidelity of ConfLab can assist in the development of more robust methods for these key tasks. We hope that models trained on ConfLab for localizing keypoints would fill the gap in the cue extraction pipeline, enabling past datasets [9, 10] without articulated pose data to be reinvigorated; this would open the floodgates for more robust analysis of the social phenomena labeled in these other datasets. Finally, our baseline social tasks form the basis for further explorations into downstream prediction tasks of socially-related constructs such as conversation quality [68], dominance [52], rapport [49], influence [69] etc.

ConfLab: the Data-Collection Concept To relate an individual's behaviors to trends within their social network, further iterations of ConfLab are needed. These iterations would enable the study of behavioral patterns at different timescales, including multiple interactions in one day, multiple days at a conference, or across distinct conferences. This paper serves as a template for such future ventures. We hope that if the idea of a conference

as a living lab gains traction, the effort and cost of data collection can be amortized across different research groups, even involving support from the conference organizers. This *data by the community for the community* ethos can enable the generation of a corpus of related datasets enabling new research questions.

Societal Impact ConfLab's long-term vision is towards developing technology to assist individuals in navigating social interactions. In this work we have identified choices that maximize data fidelity while upholding ethical best practices: an overhead camera perspective that mitigates identifying faces, recording audio at a low-frequency, and using non-intrusive wearable sensors matching a conference badge form-factor. We argue this is an essential step towards a long-term goal of developing personalized and socially aware technologies that enhance social experiences. At the same time, such interventions could also affect a community in unintended ways: worsened social satisfaction, lack of agency, stereotyping; or benefit only those members of the community who make use of resulting applications at the expense of the rest. More nefarious uses involve exploiting the data for developing methods that harmfully surveil or profile people. Researchers must consider such inadvertent effects while developing downstream applications. Finally, since we recorded the dataset at a scientific conference and required voluntary participation, there is an implicit selection bias in the population represented in the data. Researchers should be aware that insights resulting from the data may not generalize to the general population.

Empowering Users Through an Agentist Rather Than Structurist Approach The analysis of human behavior in social settings has classically taken a more top-down perspective. For instance, the analysis of situated interactions (via only proximity networks) has provided insight into the process of making science in the field of Meta Science [70]. However, while social network science is a well-populated domain, it lacks a more individualized measurement of social behavior: see more discussion of the structure vs. agency debate [71]. Relying on the network science approach jeopardizes an individual's right to technologies that enable free will. We consider the agency in choosing such technologies to be a form of individual harm avoidance. ConfLab provides access to more than just proximity data about social interactions, enabling the study of context-specific social dynamics. These dynamics are a uniquely dependent not only on the individual, but also the group they are interacting with [72]. We hope our highlighting of participatory design practices and these value-sensitive design principles promote social safety in developing socially assistive technologies.

Acknowledgements

The authors would like to thank: the ACM Multimedia 2019 General Chairs Martha Larson, Benoit Huet, and Laurent Amsaleg for their support in making the data collection at

a major international conference a reality; Bernd Dudzik, Yeshwanth Napolean, Ruud de Jong, and the venue support staff for their help in setting up the recording on site; Ioannis Protonotarios for the development of the MINGLE Midge badge; Jerry de Vos for improving our Midge Github repository and designing a new case; the participants and student volunteers for the *Meet the Chairs!* event; the Amazon Mechanical Turk workers for their efforts in annotating the dataset; Rich Radke, Martin Atzmueller, Laura Cabrera-Quiros, Alan Hanjalic, and Xucong Zhang for the insightful discussions; Santosh Ilamparuthi for the innumerable discussions and support towards strengthening the ethical soundness of recording and sharing ConfLab; Jan van der Heul for the incredibly responsive support in setting up the 4TU Data repository for ConfLab; and Bart Vastenhouw, Myrthe Tielman, and Catharine Oertel for help with the data sharing; and Musy Ayoub for the word-intelligibility analysis of the low frequency audio.

ConfLab was partially funded by Netherlands Organization for Scientific Research (NWO) under project number 639.022.606 with associated Aspasia Grant, and also by the ACM Multimedia 2019 conference via student helpers, and crane hiring for camera mounting.

References

[1] B. Dudzik, S. Columbus, T. M. Hrkalovic, D. Balliet, and H. Hung. Recognizing perceived interdependence in face-to-face negotiations through multimodal analysis of nonverbal behavior. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 121–130, 2021.

[2] W. Fleeson. Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of personality*, 75:825–862, 8 2007. doi: 10.1111/J.1467-6494.2007.00458.X. URL https://pubmed.ncbi.nlm.nih.gov/17576360/.

[3] J. G. L. Guardia and R. M. Ryan. Why identities fluctuate: Variability in traits as a function of situational variations in autonomy support. *Journal of Personality*, 75:1205–1228, 12 2007. doi: 10.1111/j.1467-6494.2007.00473.x.

[4] J. A. Hall, T. G. Horgan, and N. A. Murphy. Nonverbal communication. *Annual Review of Psychology*, 70:271–294, 1 2019. doi: 10.1146/ANNUREV-PSYCH-010418-103145. URL https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010418-103145.

[5] K. Osborne-Crowley. Social cognition in the real world: reconnecting the study of social cognition with social reality. *Review of general psychology*, 24(2):144–158, 2020.

[6] C. Andrade. Internal, external, and ecological validity in research design, conduct, and evaluation. *Indian journal of psychological medicine*, 40(5):498–499, 2018.

[7] É. Labonte-LeMoyne, F. Courtemanche, M. Fredette, and P.-M. Léger. How wild is too wild: Lessons learned and recommendations for ecological validity in physiological computing re-

search. In *PhyCS*, pages 123–130, 2018.

[8] H. Hung, E. Gedik, and L. C. Quiros. Complex conversational scene analysis using wearable sensors. In *Multimodal Behavior Analysis in the Wild*, pages 225–245. Elsevier, 2019.

[9] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung. The matchnmingle dataset: A novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, 12(1):113–130, 2021.

[10] H. Hung and B. Kröse. Detecting f-formations as dominant sets. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 231–238, 2011.

[11] X. Alameda-Pineda, J. Staiano, R. Subramanian, et al. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8): 1707–1720, 2015.

[12] M. Cristani, L. Bazzani, G. Paggetti, et al. Social interaction discovery by statistical analysis of f-formations. In *British Machine Vision Conference, BMVC 2011, Dundee, UK, August 29 - September 2, 2011. Proceedings*, pages 1–12. BMVA Press, 2011. doi: 10.5244/C.25.23. URL https://doi.org/10.5244/C.25.23.

[13] G. Zen, B. Lepri, E. Ricci, and O. Lanz. Space speaks: towards socially and personality aware visual surveillance. In *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*, pages 37–42, 2010.

[14] M. Murgia. Who's using your face? the ugly truth about facial recognition. *Financial Times*, 2019.

[15] N. Carissimi, P. Rota, C. Beyan, and V. Murino. Filling the gaps: Predicting missing joints of human poses using denoising autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[16] E. Gedik and H. Hung. Detecting conversing groups using social dynamics from wearable acceleration: Group size awareness. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(4), dec 2018. doi: 10.1145/3287041. URL https://doi.org/10.1145/3287041.

[17] E. Gedik and H. Hung. Personalised models for speech detection from body movements using transductive parameter transfer. *Personal and Ubiquitous Computing*, 21(4):723–737, Aug. 2017. doi: 10.1007/s00779-017-1006-4.

[18] C. Raman, S. Tan, and H. Hung. A modular approach for synchronized wireless multimodal multisensor data acquisition in highly dynamic social settings. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 3586–3594, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3394171.3413697. URL https://doi.org/10.1145/3394171.3413697.

[19] A. Birhane and V. U. Prabhu. Large image datasets: A pyrrhic win for computer vision? In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA,*

January 3-8, 2021, pages 1536–1546. IEEE, 2021. doi: 10.1109/WACV48630.2021.00158. URL https://doi.org/10.1109/WACV48630.2021.00158.

[20] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently Scaling up Crowdsourced Video Annotation. *International Journal of Computer Vision*, 101(1):184–204, 2013. doi: 10.1007/s11263-012-0564-1.

[21] F. Setti, C. Russell, C. Bassetti, and M. Cristani. F-formation detection: Individuating free-standing conversational groups in images. *PloS one*, 10(5):e0123783, 2015.

[22] M. Swofford, J. Peruzzi, N. Tsoi, et al. Improving social awareness through dante: Deep affinity network for clustering conversational interactants. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23, 2020.

[23] S. Vascon, E. Z. Mequanint, M. Cristani, et al. A game-theoretic probabilistic approach for detecting conversational groups. In *Asian conference on computer vision*, pages 658–675. Springer, 2014.

[24] H. Joo, T. Simon, X. Li, et al. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[25] E. Ricci, J. Varadarajan, R. Subramanian, et al. Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4660–4668, 2015.

[26] L. Bazzani, M. Cristani, D. Tosato, et al. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 30(2):115–127, 2013.

[27] J. Carletta, S. Ashby, S. Bourban, et al. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[28] Y. Cai, Z. Wang, Z. Luo, et al. Learning delicate local representations for multi-person pose estimation. In *European Conference on Computer Vision*, pages 455–472. Springer, 2020.

[29] C. Cattuto, W. V. D. Broeck, A. Barrat, et al. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLoS ONE*, 5, 2010.

[30] M. Hoffman, P. Block, T. Elmer, and C. Stadtfeld. A model for the dynamics of face-to-face interactions in social groups. *Network Science*, 8(S1):S4–S25, 2020. doi: 10.1017/nws.2020.3.

[31] M. Atzmueller and F. Lemmerich. Homophily at academic conferences. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 109–110, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[32] D. O. Olguín, B. N. Waber, T. Kim, et al. Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):43–55, 2008.

[33] D. Chaffin, R. Heidl, J. R. Hollenbeck, et al. The promise and perils of wearable sensors in organizational research. *Organizational Research Methods*, 20(1):3–31, 2017.

[34] A. Rosatelli, E. Gedik, and H. Hung. Detecting f-formations roles in crowded social scenes with wearables: Combining proxemics dynamics using lstms. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 147–153, 2019. doi: 10.1109/ACIIW.2019.8925179.

[35] L. Cabrera-Quiros, D. M.J. Tax, and H. Hung. Gestures in-the-wild : Detecting conversational hand gestures in crowded scenes using a multimodal fusion of bags of video trajectories and body worn acceleration. pages 1–10, 2018.

[36] J. V. Quiros and H. Hung. CNNs and Fisher Vectors for No-Audio Multimodal Speech Detection. In *MediaEval*, 2019.

[37] S. Tan, D. M. J. Tax, and H. Hung. Multimodal joint head orientation estimation in interacting groups via proxemics and interaction dynamics. *Proc. ACM Interactive, Mobile, Wearable, and Ubiquitous Technology*, 5(1), Mar. 2021.

[38] University of york research data management. `https://www.york.ac.uk/library/info-for/researchers/data/sharing/access/`.

[39] Utrecht university research data management. `https://www.uu.nl/en/research/research-data-management/guides/handling-personal-data`.

[40] Go pro hero 7 black. `https://gopro.com/en/nl/shop/cameras/hero7-black/CHDHX-701-master.html`.

[41] O. Lederman, A. Mohan, D. Calacci, and A. S. Pentland. Rhythm: A unified measurement platform for human organizations. *IEEE MultiMedia*, 25(1):26–38, 2018.

[42] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013.

[43] L. Cabrera-Quiros and H. Hung. Who is where? matching people in video to wearable acceleration during crowded mingling events. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 267–271, 2016.

[44] L. Cabrera-Quiros and H. Hung. A hierarchical approach for associating body-worn sensors to video regions in crowded mingling scenarios. *IEEE Transactions on Multimedia*, 21(7):1867–1879, 2018.

[45] H. Hung, C. Raman, E. Gedik, S. Tan, and J. Vargas Quiros. Multimodal data collection for social interaction analysis in-the-wild. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2714–2715, 2019.

[46] Computer Vision Annotation Tool (CVAT).

[47] J. Vargas Quiros, S. Tan, C. Raman, L. Cabrera-Quiros, and H. Hung. Covfee: an extensible web framework for continuous-time annotation of human behavior. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, volume 173 of *Proceedings of Machine Learning Research*, pages 265–293. PMLR, 16 Oct 2022. URL `https://proceedings.mlr.press/v173/vargas-quiros22a.html`.

[48] D. Gatica-Perez. Analyzing group interactions in conversations: a review. In *2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 41–46, 2006. doi: 10.1109/MFI.2006.265658.

[49] P. Müller, M. X. Huang, and A. Bulling. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In *23rd International Conference on Intelligent User Interfaces*. ACM, 2018. doi: 10.1145/3172944.3172969.

[50] C. Raman and H. Hung. Towards automatic estimation of conversation floors within F-formations. *arXiv:1907.10384 [cs]*, July 2019.

[51] H. Hung and D. Gatica-Perez. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia*, 12(6):563–575, 2010.

[52] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez. Estimating Dominance in Multi-Party Meetings Using Speaker Diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):847–860, may 2011.

[53] C. Beyan, M. Shahid, and V. Murino. RealVAD: A Real-world Dataset and A Method for Voice Activity Detection by Body Motion Analysis. 9210:1–16, 2020. doi: 10.1109/tmm.2020.3007350.

[54] M. Shahid, C. Beyan, and V. Murino. Voice activity detection by upper body motion analysis and unsupervised domain adaptation. *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*, pages 1260–1269, 2019. doi: 10.1109/ICCVW.2019.00159.

[55] Covfee: Continuous Video Feedback Tool. Jose Vargas.

[56] A. Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[57] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[58] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[59] T.-Y. Lin, M. Maire, S. Belongie, et al. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[60] J. Shen, O. Lederman, J. Cao, et al. GINA: Group Gender Identification Using Privacy-Sensitive Audio Data. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2018-Novem: 457–466, 2018. doi: 10.1109/ICDM.2018.00061.

[61] P. Gupta, A. Thatipelli, A. Aggarwal, et al. Quo Vadis, Skeleton Action Recognition ? *arXiv:2007.02072 [cs]*, July 2020.

[62] Z. Wang, W. Yan, and T. Oates. Time series classification from scratch with deep neural networks: A strong baseline, 2016. URL `https://arxiv.org/abs/1611.06455`.

[63] Z. Wang, W. Yan, and T. Oates. Time series classification from scratch with deep neural networks: A strong baseline. *Proceedings of the International Joint Conference on Neural Networks*, 2017-May:

1578–1585, 2017. doi: 10.1109/IJCNN.2017.7966039.

[64] C. W. Tan, A. Dempster, C. Bergmeir, and G. I. Webb. Multirocket: Multiple pooling operators and transformations for fast and effective time series classification, 2021. URL https://arxiv.org/abs/2102.00457.

[65] T. Baltrusaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, Feb. 2019.

[66] J. Liu, A. Shahroudy, M. Perez, et al. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10): 2684–2701, Oct. 2020. doi: 10.1109/TPAMI.2019.2916873.

[67] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino. Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, 100:86–97, 2013.

[68] C. Raman, N. R. Prabhu, and H. Hung. Perceived conversation quality in spontaneous interactions, 2022. URL https://arxiv.org/abs/2207.05791.

[69] W. Dong, B. Lepri, A. Cappelletti, et al. Using the influence model to recognize functional roles in meetings. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 271–278, 2007.

[70] J. Eberle, K. Stegmann, A. Barrat, F. Fischer, and K. Lund. Initiating scientific collaborations across career levels and disciplines–a network analysis on behavioral data. *International Journal of Computer-Supported Collaborative Learning*, 16(2):151–184, 2021.

[71] N. Pleasants. Free will, determinism and the "problem" of structure and agency in the social sciences. *Philosophy of the Social Sciences*, 49(1):3–30, 2019.

[72] C. Raman, H. Hung, and M. Loog. Social Processes: Self-Supervised Meta-Learning over Conversational Groups for Forecasting Nonverbal Social Cues. *arXiv:2107.13576 [cs]*, July 2021.

[73] T. Gebru, J. Morgenstern, B. Vecchione, et al. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[74] G. Barquero, J. Núnez, S. Escalera, et al. Didn't see that coming: a survey on non-verbal social human behavior forecasting. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 139–178. PMLR, 2022.

[75] C. Raman, H. Hung, and M. Loog. Why did this model forecast this future? closed-form temporal saliency towards causal explanations of probabilistic forecasts. *arXiv preprint arXiv:2206.00679*, 2022.

[76] H. Hung, G. Englebienne, and J. Kools. Classifying social actions with a single accelerometer. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 207–210, 2013.

[77] N. Raj Prabhu, C. Raman, and H. Hung. Defining and quantifying conversation quality in spontaneous interactions. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 196–205, 2020.

[78] J. D. V. Quiros, O. Kapcak, H. Hung, and L. Cabrera-Quiros. Individual and joint body movement assessed by wearable sensing as a predictor of attraction in speed dates. *IEEE Transactions on Affective Computing*, 2021.

[79] OpenCV. Open source computer vision library. https://github.com/opencv/opencv, 2015.

[80] Idiap multi camera calibration suite. https://github.com/idiap/multicamera-calibration.

[81] Tdkicm20948. https://invensense.tdk.com/products/motion-tracking/9-axis/icm-20948/. Accessed: 2021-10-15.

[82] S. O. Ba and J.-M. Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):101–116, 2010.

[83] W. Li, Z. Wang, B. Yin, et al. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019.

[84] B. Cheng, B. Xiao, J. Wang, et al. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[85] A. Newell, Z. Huang, and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017.

[86] I. Oguiza. tsai - a state-of-the-art deep learning library for time series and sequential data. Github, 2022. URL https://github.com/timeseriesAI/tsai.

Appendices

2.A Hosting, Licensing, and Organization

The dataset is hosted by 4TU.ResearchData, available at https://doi.org/10.4121/c.6034313.

The dataset itself is available under restricted access defined by an End-User License Agreement (EULA). The EULA itself is available under a CC0 license. The code (https://github.com/TUDelft-SPC-Lab/conflab) for the benchmark baseline tasks, and the schematics and data associated with the design of our custom wearable sensor called the Midge (https://github.com/TUDelft-SPC-Lab/spcl_midge_hardware) are available under the MIT License.

Figure 2.10 on the next page illustrates the organization of the ConfLab dataset on 4TU.ResearchData. The components are as follows:

- Annotations (restricted, https://doi.org/10.4121/20017664): annotations of pose, speaking status, and F-formations
- Datasheet for ConfLab (public, https://doi.org/10.4121/20017559): documentation of the dataset following Datasheets for Datasets [73] (see Appendix 2.B)
- EULA (public, https://doi.org/10.4121/20016194): End User License Agreement to be signed for requesting access to the restricted components
- Processed-Data (restricted, https://doi.org/10.4121/20017805): processed video and wearable sensor used for annotations
- Raw-Data (restricted, https://doi.org/10.4121/20017748): raw video and wearable sensor data
- Data Samples (restricted, https://doi.org/10.4121/20017682): samples of the sensor, audio, and video data

```
ConfLab/
├── Annotations/
│   ├── actions/
│   │   ├── confidence/
│   │   │   └── <video_seg>.csv
│   │   ├── raw-covfee.zip/
│   │   │   └── <video_seg>.json
│   │   └── speaking/
│   │       └── <video_seg>.csv
│   ├── f_formations/
│   │   ├── <video_seg>.csv
│   │   └── phone.csv
│   └── pose/
│       ├── coco/
│       │   └── <video_seg>.json
│       └── raw-covfee.zip/
│           └── <video_seg>.json
├── datasheet.pdf
├── EULA.pdf
├── Processed-Data/
│   ├── cameras/
│   │   ├── camera-calibration/
│   │   │   └── <cam_num>/
│   │   │       └── intrinsics.json
│   │   └── video_segments/
│   │       └── <video_seg>.MP4
│   └── wearables/
│       └── <person_id>/
│           ├── <person_id>_accel.csv
│           ├── <person_id>_audio
│           ├── <person_id>_gyro.csv
│           ├── <person_id>_mag.csv
│           ├── <person_id>_proximity.csv
│           └── <person_id>_rotation.csv
├── Raw-Data/
│   ├── audio/
│   │   ├── sync_files/
│   │   │   └── auxiliary_sync.csv
│   │   └── synced/
│   │       └── <person_id>_audio.wav
│   ├── cameras/
│   │   ├── camera-calibration/
│   │   │   └── <cam_num>/
│   │   │       └── intrinsics.json
│   │   └── video/
│   │       └── <cam_num>/
│   │           └── <video_seg_from_go_pro>.MP4
│   └── wearables/
│       └── <person_id>/
│           ├── <person_id>_accel.csv
│           ├── <person_id>_audio
│           ├── <person_id>_gyro.csv
│           ├── <person_id>_mag.csv
│           ├── <person_id>_proximity.csv
│           └── <person_id>_rotation.csv
└── Samples/
    ├── sample_audio
    ├── sample_audio.wav
    ├── sample_sensor.csv
    └── sample_video.MP4
```

Figure 2.10: File structure of the ConfLab dataset

2.B Datasheet For ConfLab

This document is based on *Datasheets for Datasets* by Gebru *et al.* [73]. Please see the most updated version here.

MOTIVATION

Q. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

There are two broad motivations for creating this dataset: first, to enable the privacy-preserving, multimodal study of *real-life* social conversation dynamics; second, to bring the higher fidelity of wired in-the-lab recording setups to in-the-wild scenarios, enabling the study of *fine time-scale* social dynamics in-the-wild.

We propose the Conference Living Lab (ConfLab) with the following goals: (i) a data collection effort that follows a *by the community for the community* ethos: the more volunteers, the more data, (ii) volunteers who potentially use the data can experience first-hand potential privacy and ethical considerations related to sharing their own data, (iii) in light of recent data sourcing issues [19], we incorporated privacy and invasiveness considerations directly into the decision-making process regarding sensor type, positioning, and sample-rates.

From a technical perspective, closest related datasets (see Table 2.1 in the main paper) suffer from several technical limitations precluding the analysis and modeling of fine-grained social behavior: (i) lack of articulated pose annotations; (ii) a limited number of people in the scene, preventing complex interactions such as group splitting/merging behaviors, and (iii) an inadequate data sampling-rate and synchronization-latency to study time-sensitive social phenomena [18, Sec. 3.3]. This often requires modeling simplifications such as the summarizing of features over rolling windows [17, 35, 36]. On the other hand, past high-fidelity datasets have largely involved role-played or scripted interactions in lab settings, with often a single-group in the scene.

This dataset wasn't created with a specific task in mind, but intends to support a wide variety of multimodal modeling and analysis tasks across research domains (see the *Uses* section).

Q. Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

ConfLab was initiated by the Socially Perceptive Computing Lab, Delft University of Technology in cooperation and support from the general chairs of ACM Multimedia 2019 (Martha Larson, Benoit Huet, and Laurent Amsaleg), Nice, France. Since this dataset was by the community, for the community, members of the Multimedia community contributed

as subjects in the dataset.

Q. What support was needed to make this dataset? (e.g. who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

ConfLab was partially funded by Netherlands Organization for Scientific Research (NWO) under project number 639.022.606 with associated Aspasia Grant, and also by the ACM Multimedia 2019 conference via student helpers, and crane hiring for camera mounting.

Q. Any other comments?
None.

COMPOSITION

Q. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The dataset contains multimodal recordings of people interacting during a networking event embedded in an international multimodal machine learning conference. Overall, the interaction scene contained conversation groups (operationalized as f-formations), composed of individual subjects, each of which had individual data associated to their wearable sensors. The complete interaction scene was additionally captured by overhead cameras. Figure 2.11 shows the structure of these instances and their relationships.

Note however that the precise notion of what constitutes an instance in the dataset is very much task-specific. In our baseline tasks we considered the following instances:

Person and Keypoints Detection Frames, containing pose annotations (17 body keypoints per person per frame @60 Hz) from 5 overhead videos ($1920 \times 1080, 60$ fps) for 16 minutes of interaction.

Speaking Status Detection Windows (3 seconds) of wearable sensor data and speaking status annotations (60 Hz) extracted from each subject's data.

F-formations Operationalized conversation groups, annotated at 1 Hz from the 16 minutes of annotated data, and the pose data associated to the people in the F-formation.

Q. How many instances are there in total (of each type, if appropriate)?
The notion of instance is very much dependent on how a user intends to use the data. Regarding the instances in Figure 2.11, our full dataset consist of 45 minutes of:

Video recordings from 10 overhead cameras placed over the interaction area. Five of these videos, enough to cover the complete interaction area, were used in annotation.

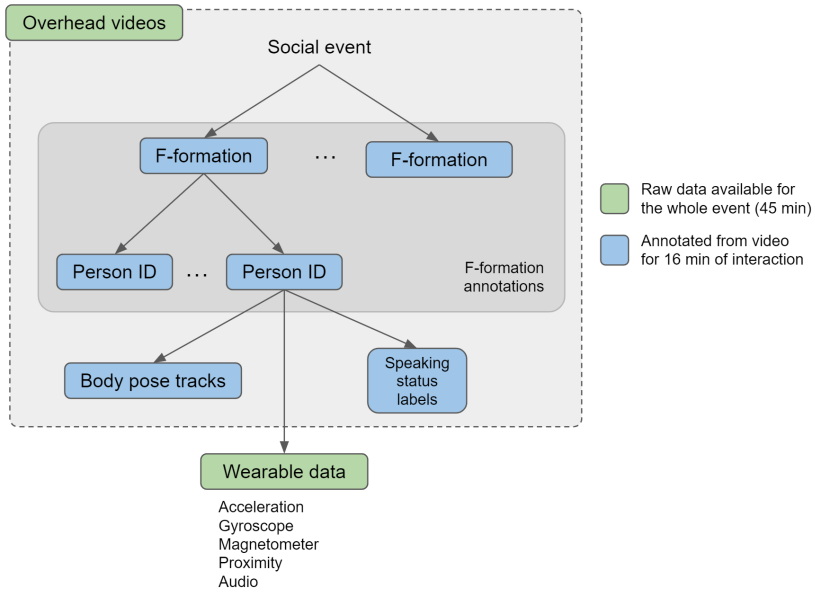


Figure 2.11: Structure of some of the instances in the dataset and their relationships. The interaction space was captured via overhead videos, in which f-formations (conversation groups) were annotated. An F-formation consists of set of people interacting for a variable period of time, and identified via a subject ID. Each person in the F-formation can be associated to their pose (annotated in the videos), their wearable sensor (IMU) data, and their action (speaking status) labels.

Individual wearable sensor data For the 48 subjects in the interaction area, a chest-worn conference-type badge recorded: audio (1250 Hz), and Inertial Measurement Unit (IMU) readings (accelerometer @ 56 Hz, gyroscope @56 Hz, magnetometer @56 Hz and Bluetooth RSSI-based proximity @5 Hz)

Conference experience label For each of the 48 subjects, an associated self-report label indicating whether it was their first time in the conference.

The instances in the annotated 16 minutes segment out of the 45 minutes of interaction contain:

2D body poses For each of the 48 subjects, full body pose tracks annotated at 60Hz (17 keypoints per person). These were annotated using 5 of the 10 overhead cameras due to the significant overlap in views (cameras 2, 4, 6, 8, and 10). Annotations were done separately for each camera by annotating all of the people visible in each video, for each of the 5 cameras, and tagged with a participant ID. We made use of a novel continuous technique for annotation of keypoints. We chose this approach via a pilot study with 3 annotators, comparing our technique to annotations done using the non-continuous CVAT tool. We found no statistically significant differences in

errors per-frame (as measured using Mean Squared Error across annotators), despite a 3x speed-up in annotation time in the continuous condition. The details of the technique and this pilot study can be found in [47].

Speaking status annotations For each of the 48 subjects, these include a) a binary signal (60 Hz) indicating whether the person is perceived to be speaking or not; b) continuous confidence value (60 Hz) indicating the degree of confidence of the annotator in their speaking status assessment. These annotations were done without access to audio due to issues with the synchronization of the audio recordings at the time of annotation. The confidence assessment is therefore largely based on the visibility of the target person and their speaking-associated gestures (eg. occlusion, orientation w.r.t. camera, visibility of the face)? We measured inter-annotator agreement for speaking status in a pilot where two annotators labeled three data subjects for 2 minutes each. We measured a frame-level agreement (Fleiss' κ) of 0.552, comparable to previous work [35].

F-formation annotations These annotations label the conversing groups in the scene following previous work. Each individual belongs to one F-formation at a time or is a singleton in the interaction scene. The membership is binary. The annotations were done by one of the authors at 1 Hz by watching the video. The time-stamped usage of mobile phones are available as auxiliary annotations, which are useful for the study of the role of mobile phone users as associates of F-formations. Since Kendon's theories date back to before the widespread use of mobile phones, their influence on F-formation membership remains an open question.

In our baseline tasks, which made use of the complete annotated section of the dataset, the instance numbers were the following:

Person and Keypoints Detection 119k frames (60fps) containing 1967k person instances (poses) in total, from 48 subjects recorded in 5 cameras (16 minutes of annotated segment).

Speaking Status Detection 42884 3-second windows, extracted from the 48 participants' wearable data and speaking status annotations.

F-formations 119 conversation groups. Details are in Section 2.5.

Q. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The participants in our data collection are a sample of the conference attendees. Participants were recruited via the conference website, social media posting, and approaching them in

person during the conference. Because participation in such a data collection can only be voluntary, the sample was not pre-designed and may not be representative of the larger set.

Additionally, 16 minutes of sensor data has been annotated for keypoints, speaking status and F-formations out of the total of 45 minutes recorded. The remaining part (across all modalities) is provided with no labels. For privacy reasons, the elevated cameras (distinct from the previously mentioned 8 overhead cameras) and also individual frontal headshots that were used for manually associating the video data to the wearable sensor data is not being shared.

Q. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Camera 5 failed early during the recording, but the space underneath it was captured by the adjacent cameras due to the high overlap in the camera field-of-views. Nevertheless we share what was recorded before the failure from camera 5, bringing the total number of cameras to 9.

Q. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

The F-formations, subjects, and their associated data relate as shown in Figure 2.11. These associations are made explicit in the dataset via anonymous subject IDs, associated to pose tracks, speaking status annotations, and wearable sensor data. These same IDs were used to annotate the F-formations.

Pre-existing personal relationships between the subjects were not requested for privacy reasons.

Q. Are there recommended data splits (e.g., training, development/validation, testing)?

Since the dataset can be used to study a variety of tasks, the answer to this question is task dependent. Please refer to our reproducibility details (Appendix 2.G of our associated paper) for information about the splits that we used in out baselines.

Q. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Individual audio Because audio was recorded by a front-facing wearable device worn on the chest, it contains a significant amount of cocktail party noise and cross-contamination from other people in the scene. In our experience this means that automatic speaking status detection is challenging with existing algorithms but manual annotation is possible.

Videos and 2D body poses It is important to consider that the same person may appear in multiple videos at the same time if the person was in view of multiple cameras. Because 2D poses were annotated per video, the same is true of pose annotations. Each skeleton was tagged with a person ID, which should serve to identify such cases when necessary.

Q. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained.

Q. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?

The data contains personal data under GDPR in the form of video and audio recordings of subjects. The dataset is shared under an End User License Agreement for research purposes, to ensure that the data is not made public, and to protect the privacy of data subjects.

Q. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No.

Q. Does the dataset relate to people?

Yes, the dataset contains recordings of human subjects.

Q. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Data subjects answered the following questions before the start of the data collection event, after filling in their consent form:

- Is this your first time attending ACM MM?
- Select the area(s) that describes best your research interest(s) in recent years. Descriptions of each theme are listed here: https://acmmm.org/call-for-papers/

Figure 2.12 shows the distribution of the responses / populations.

Q. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

We do not share any directly identifiable information as part of the dataset. However, individuals may be identified in the video recordings if the observer knows the participants in the recordings personally. Otherwise, individuals in the dataset may potentially be identified in combination with publicly available pictures or videos (from conference attendees or conference official photographer) from other media from the conference the dataset was recorded at. In any case, re-identifying the subjects is strictly against the End User License Agreement under which we share the dataset.

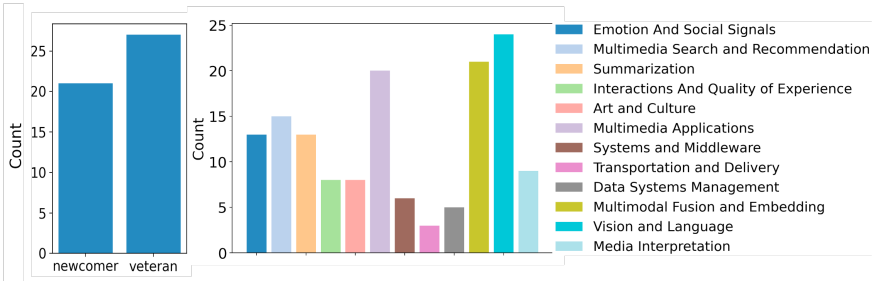


Figure 2.12: Distribution of participant seniority (left) and research interests (right) in percentage.

Q. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?
We did not request any such information from data participants. Here, the ACM Multimedia '19 General Chair Martha Larson also helped advocate on behalf of the attendees during the survey-design stage. As a result of these discussions, information such as participant gender, ethnicity, or country of origin was not asked.
Q. Any other comments?
None.

COLLECTION

Q. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The collected data is directly observable, containing video recordings, low-frequency audio recordings and wearable sensing signals (inertial motion unit (IMU) and Bluetooth proximity sensors) of individuals in the interaction scenes. Accompanying data includes self-reported binary categorization of experience level which is available upon request from the authors. The self-reported interests categories are not shared because of privacy concerns.
Video recordings capture the whole interaction floor where the association from multimodal data to individual is done manually by annotators by referring to frontal (not-shared) and overhead views. The rest of the data was acquired from the wearable sensing badges,

which is person-specific (i.e., no participant shared the device). Video and audio data were verified in playback. Wearable sensing data was verified through plots after parsing.

Q. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the s was created. Finally, list when the dataset was first published.

All data was collected on October 24, 2019, except the self-reported experience level and research interest topics which are either obtained on the same day or not more than one week before the data collection day. This time frame matches the creation time frame of the data association for wearable sensing data. Video data was associated with individual during annotation stage (2020-2021), but all information used for association was obtained on the data collection day.

Q. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

To record videos, we used 14 GoPro Hero 7 Black cameras. The wearable sensor hardware has been documented and open-sourced at https://github.com/TUDelft-SPC-Lab/spcl_midge_hardware. The validation of the sensors was completed through an external contractor engineer. The data collection software was documented and published in [47], which includes validation of the system. These hardwares and mechanisms have been open-sourced along with their respective publication.

The synchronization setup for data collection (intramodal and intermodal) was documented and published in [18], which includes validation of the system.

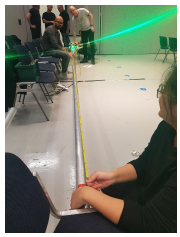
To lend the reader further insight into the process of setting up the recording of such datasets in-the-wild, we share images of our process in Figure 2.13.

Q. What was the resource cost of collecting the data?

The resources required to run this first edition of ConfLab include equipment, logistics, and travel costs. Table 2.5 shows the full breakdown of the costs. The equipment expenses are fixed one-time costs since the same equipment can be used for future iterations of ConfLab. The on-site costs at the conference venue were toward renting a crane for a day to mount the cameras on a scaffold on the ceiling. We have open-sourced the Midge (our custom wearable) schematics so that others don't need to spend on the design and development.

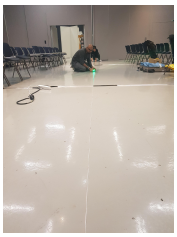
No additional energy consumption was incurred for collecting the data. However, the ancillary activities (e.g., flights, accommodation) resulted in energy consumption. Flights from the Netherlands to France round-trip for six passengers results in 1020 kg carbon emissions. Accommodation for six members resulted in 22 kWh energy consumption.

Q. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?



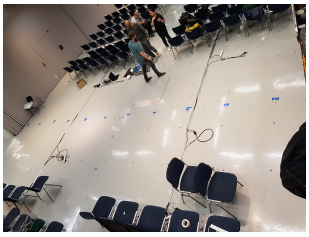
(a) Aligning cameras

(b) Affixing the mounting beam



(c) Aligning floor markers

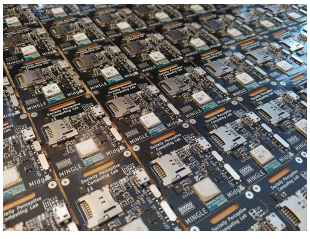
(d) Marking the floor grid



(e) Interaction area

(f) Verifying camera sync.

(g) Assembling Midges



(h) Midges

(i) Verifying crossmodal sync.

Figure 2.13: Illustrating the process of setting up the data recording.

ConfLab contains both annotated and unannotated segments of multi-modal data. The segment where the articulated pose and speaking status were annotated is selected to maximize crowd density in the scenes. The annotated segment is 16 minutes; the whole set is roughly 1 hour of recordings.

Q. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The Conflab dataset was captured during a special social event called *Meet the Chairs!* at an international conference on signal processing and machine learning. Newcomers and old-timers to the conference freely donated their social behaviour data as part of a *by the community, for the community* data collection effort. Aside from the chance to meet the chairs and create a community dataset, the attendees also received a personalised report of their social behaviour from the wearable sensors (see Appendix 2.C) Conference

Table 2.5: Itemized costs associated with recording ConfLab

Item	Cost (USD)
Travel (total for 6 people)	
Flights	1800
Accommodation	1500
Equipment (one time)	
Mounting scaffold	2000
14 × GoPro Hero 7 Black	4900
Designing the Midge (custom wearable, now made open source)	26000
110 × Midges (boards, batteries, 4 GB sd cards, cases)	3660
Multimodal synchronization setup	730
Annotations	8000
Computational cost for experiments	500

student volunteers were involved in assisting the set-up of the event. Conference organizers (mentioned in the *Motivation* section) assisted in connecting us with conference venue contacts to mount our technical set-ups in the room. Volunteers and conference organizers were not paid by us. Conference venue contacts were paid by the conference organizers.

Data annotations were completed by crowdsourced workers. The crowdsourced workers were paid $0.20 for qualification assignment (note that typically requesters do not pay for qualification tasks). Depending on the submitted results, workers earn qualification to access of the actual tasks. The annotation tasks were categorized into low-effort ($150), medium-effort ($300), and high-effort ($450), corresponding to the amount of estimated time each would take. The duration of the tasks was determined by the crowd density and through timing of the pilot studies. The average hourly payment to workers is around $8.

Q. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

The data collection was approved by the Human Research Ethics Committee (HREC) of our university (Delft University of Technology), which reviews all research involving human subjects. The data collection protocol is also compliant to the conference location's national authorities (France). The review process included addressing privacy concerns to ensure compliance with GDPR and university guidelines, review of our informed consent form, data management plan, and end user license agreement for the dataset and a safety check of our custom wearable devices.

Q. Does the dataset relate to people?

Yes.

Q. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

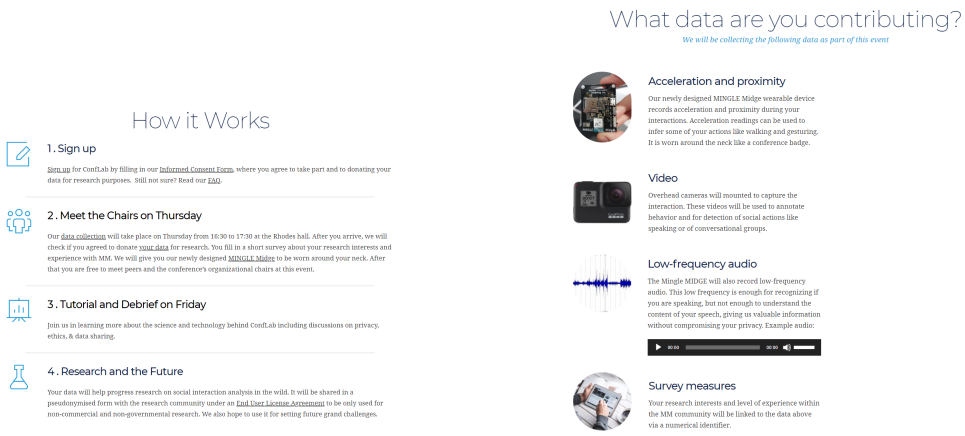


Figure 2.14: Screenshots of the ConfLab web-page used for participant recruitment and registration.

We collected the data from individuals directly.

Q. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

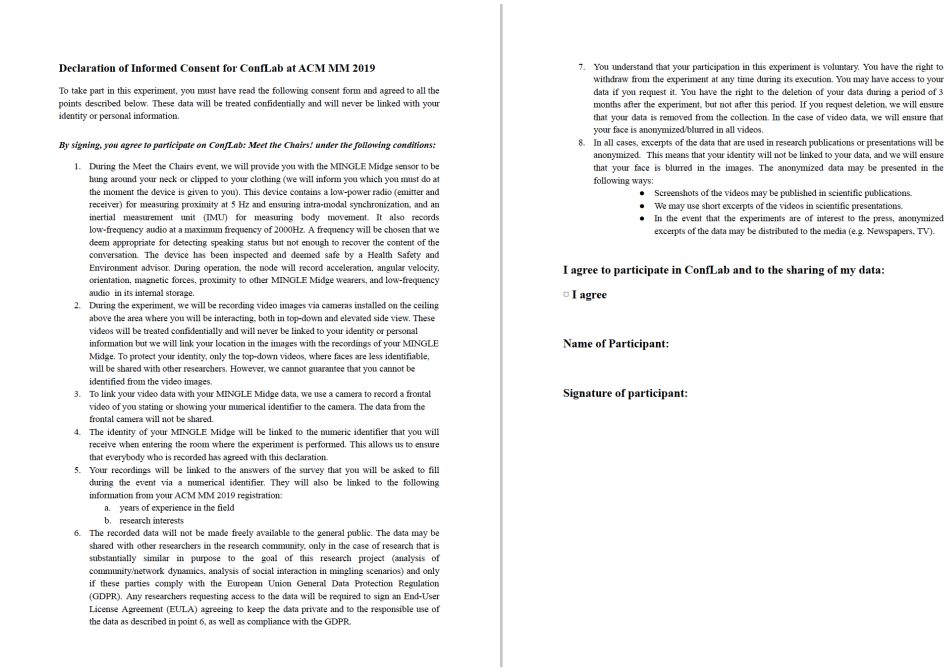
The individuals were notified about the data collection and their participation is voluntary. The data collection was staged at an event called *Meet the Chairs* at ACM MM 2019. The ConfLab web page (https://conflab.ewi.tudelft.nl/) served to communicate the aim of the event, what was being recorded, and how participants could sign up. This allowed us to embed the informed consent into this framework so we could keep track of sign ups. See Figure 2.14 for screenshots. This event website was also shared by the conference organizers and chairs (https://2019.acmmm.org/conflab-meet-the-chairs/index.html).

Q. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

All the individuals who participated in the data collection gave their consent by signing a consent form. A copy of the form is attached below in Figure 2.15.

Q. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)

Yes, the consenting individuals were informed about the possibility of revoking access to their data within a period of 3 months after the data collection experiment, and not after that. The description is included in the consent form.



Declaration of Informed Consent for ConfLab at ACM MM 2019

To take part in this experiment, you must have read the following consent form and agreed to all the points described below. These data will be treated confidentially and will never be linked with your identity or personal information.

By signing, you agree to participate on ConfLab: Meet the Chairs! under the following conditions:

1. During the Meet the Chairs event, we will provide you with the MINGLE Midge sensor to be hung around your neck or clipped to your clothing (we will inform you which you must do at the moment the device is given to you). This device contains a low-power radio (emitter and receiver) for measuring proximity at 5 Hz and ensuring intra-modal synchronization, and an inertial measurement unit (IMU) for measuring body movement. It also records low-frequency audio at a maximum frequency of 2000Hz. A frequency will be chosen that we deem appropriate for detecting speaking status but not enough to recover the content of the conversation. The device has been inspected and deemed safe by a Health Safety and Environment advisor. During operation, the node will record acceleration, angular velocity, orientation, magnetic forces, proximity to other MINGLE Midge wearers, and low-frequency audio in its internal storage.
2. During the experiment, we will be recording video images via cameras installed on the ceiling above the area where you will be interacting, both in top-down and elevated side view. These videos will be treated confidentially and will never be linked to your identity or personal information but we will link your location in the images with the recordings of your MINGLE Midge. To protect your identity, only the top-down videos, where faces are less identifiable, will be shared with other researchers. However, we cannot guarantee that you cannot be identified from the video images.
3. To link your video data with your MINGLE Midge data, we use a camera to record a frontal video of you stating or showing your numerical identifier to the camera. The data from the frontal camera will not be shared.
4. The identity of your MINGLE Midge will be linked to the numeric identifier that you will receive when entering the room where the experiment is performed. This allows us to ensure that everybody who is recorded has agreed with this declaration.
5. Your recordings will be linked to the answers of the survey that you will be asked to fill during the event via a numerical identifier. They will also be linked to the following information from your ACM MM 2019 registration:
    a. years of experience in the field
    b. research interests
6. The recorded data will not be made freely available to the general public. The data may be shared with other researchers in the research community, only in the case of research that is substantially similar in purpose to the goal of this research project (analysis of community/network dynamics, analysis of social interaction in mingling scenarios) and only if these parties comply with the European Union General Data Protection Regulation (GDPR). Any researchers requesting access to the data will be required to sign an End-User License Agreement (EULA) agreeing to keep the data private and to the responsible use of the data as described in point 6, as well as compliance with the GDPR.
7. You understand that your participation in this experiment is voluntary. You have the right to withdraw from the experiment at any time during its execution. You may have access to your data if you request it. You have the right to the deletion of your data during a period of 3 months after the experiment, but not after this period. If you request deletion, we will ensure that your data is removed from the collection. In the case of video data, we will ensure that your face is anonymized/blurred in all videos.
8. In all cases, excerpts of the data that are used in research publications or presentations will be anonymized. This means that your identity will not be linked to your data, and we will ensure that your face is blurred in the images. The anonymized data may be presented in the following ways:
    - Screenshots of the videos may be published in scientific publications.
    - We may use short excerpts of the videos in scientific presentations.
    - In the event that the experiments are of interest to the press, anonymized excerpts of the data may be distributed to the media (e.g. Newspapers, TV).

I agree to participate in ConfLab and to the sharing of my data:

□ **I agree**

Name of Participant:

Signature of participant:

Figure 2.15: Consent form signed by each participant in the data collection.

Q. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?
No.
Q. Any other comments?
None.

PREPROCESSING / CLEANING / LABELING

Q. Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
We did not pre-process the signals obtained from the wearable devices or cameras. The only exception is the audio data. Due to a hardware malfunction (this is resolved for the Midges by using different SD cards), the audio needed to be post-processed in order to synchronize it with the other modalities. The synchronization against other modalities was manually checked.
Labeling of the dataset was done as explained in the *Composition* section.

Q. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
The dataset is separated into raw data and the post processed data. For the audio, the original raw data is not suitable for most use cases due to the mentioned synchronization issue. So we share the synchronized version in the raw part of the repository.
Q. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
The processing / fixing of the audio files did not require special software.
The annotation of keypoints and speaking status was done by making use of the Covfee framework: https://josedvq.github.io/covfee/
Q. Any other comments?
None.

USES

Q. Has the dataset been used for any tasks already? If so, please provide a description.
In the main paper, we have benchmarked three baseline tasks: person and keypoints detection, speaking status detection, and F-formation detection. The first task is a fundamental building block for automatically analyzing human social behaviors. The other two demonstrate how learned body keypoints can be used in the behavior analysis pipeline for inferring more socially related phenomena. We chose these benchmarking tasks since they have been studied on other in-the-wild behavior datasets.
Q. Is there a repository that links to any or all papers or systems that use the dataset?
None at the time of writing of the paper.
Q. What (other) tasks could the dataset be used for?
Given the richness and the unscripted open-ended nature of the social interactions, ConfLab can be used for many other tasks.

Forecasting, causal relationship discovery Recently, tasks pertaining to the forecasting low-level social cues in conversations have been receiving increased attention from the community [72, 74]. The real-life nature of ConfLab along with the increased data and annotation fidelity can prove a valuable resource for such tasks. Similarly, ConfLab can also be used for efforts towards discovering causal relationships between social behaviors [75].

Data Association. A crucial assumption made in many former multimodal datasets[9, 11, 24] is that the association of video data to the wearable modality can be manually

performed. Few works [43, 44] have tried to address this issue but using movement cues alone to associate the modalities is challenging as conversing individuals are mostly stationary. This remains a significant and open question for future large scale deployable multimodal systems. One solution may be to annotate more social actions as a form of top-down supervision. However, detecting pose and actions robustly from overhead cameras remains to be solved.

Conversation floor and F-formation estimation Prior analysis on the MatchNMingle dataset has demonstrated that F-formations can contain multiple simultaneous conversations when the F-formations contain a least 4 people [50]. If this is the case for the ConfLab dataset, this may drastically change how F-formations should be labelled (e.g. returning to being a more subjective task [10]) as more time-precise labelling could enable a more nuanced take on F-formation and conversation floor membership over time.

Multi-class social action estimation More annotations resources were focused on speaker status, F-formation, and keypoint estimation. However, there are a wealth of other social actions in the data that could be interesting to combine into a more complex multi-class social action estimation task. Example social actions include drinking, mobile phone use, hand and head gesture types [9, 76].

Estimation and analysis of socially-related phenomena Beyond the modeling of human behavior which is of interest to the Computer Vision and Machine Learning communities, our benchmarked tasks form the basis for further explorations into downstream prediction of socially-related constructs which is of interest to the Social Science and Social Psychology communities. Such constructs include conversation quality [68, 77], dominance [52], rapport [49], and influence [69].

Investigation of novel crossmodal fusion strategies The baseline tasks in our paper rely only on a late fusion strategy. However, ConfLab's sub-second expected cross modal latency of ~ 13 ms along with higher sampling rate of features (60 fps video, 56 Hz IMU) opens the gateway for the in-the-wild study of nuanced time-sensitive social behaviors like mimicry and synchrony (for predicting e.g. attraction [78]) which need tolerances as low as 40 ms [18, Sec.3.2]. Prior works coped with lower tolerances by computing summary statistics over input windows [17, 35, 36]. ConfLab enables for the first time, the exploration of Multimodal machine learning approaches for social behaviour analysis in these highly dynamic in-the-wild settings [65]. Through the provided annotations Conflab also enables research in the topic of usage of mobile phones in small-group social interactions in-the-wild.

Person attribute estimation Estimating individuals that are newcomers/old timers from the dataset may be possible based on their networking strategies.

Q. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Although ConfLab's long-term vision is towards developing technology to assist individuals in navigating social interactions, the data could also affect a community in unintended ways: for instance, cause worsened social satisfaction, a lack of agency, stereotype newcomers and veterans, or benefit only those members of the community who make use of resulting applications at the expense of the rest. More nefarious uses involve exploiting the data for developing methods that harmfully surveil or profile people. Researchers must consider such inadvertent effects must while developing downstream applications. Finally, since we recorded the dataset at a scientific conference and required voluntary participation, there is an implicit selection bias in the population represented in the data. Consequently, researchers using the data should be aware that resulting insights may not generalize to the general population.

Q. Are there tasks for which the dataset should not be used? If so, please provide a description.

Beyond the cautionary discussion in the previous question, tasks involving the re-identifying the subjects is strictly against the End User License Agreement under which we share the dataset.

Q. Any other comments?

None.

DISTRIBUTION

Q. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset is available for third parties outside of Delft University of Technology to use for academic research purposes subject signing and approval of our End User License Agreement. The dataset will be hosted by 4TU.ResearchData (see the Maintenance section for description of the 4TU entity).

Q. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be distributed via the 4TU.ResearchData user interface where the data can be downloaded. The dataset has a DOI: https://doi.org/10.4121/c.6034313

Q. When will the dataset be distributed?
The dataset has been available since June 9, 2022.
Q. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
The dataset will be distributed under a restricted copyleft license, specified within our End User License Agreement, accessible through the 4TU.ResearchData dataset website. No fees are associated with the license.
Q. Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
No.
Q. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
The terms of our EULA and the European General Data Protection Regulations (GDPR) apply.
Any other comments?
None.

MAINTENANCE

Q. Who is supporting/hosting/maintaining the dataset?
The dataset is hosted by 4TU.ResearchData (https://www.4tu.nl/en/about_4tu/), and supported and maintained by The Socially Perceptive Computing Lab at TUDelft.
Q. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Via email: SPCLabDatasets-insy@tudelft.nl.
Q. Is there an erratum?
No.
Q. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
Updates will be done as needed as opposed to periodically. Instances could be deleted, added, or corrected. The updates will be posted on the 4TU.ResearchData dataset website.
Q. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

No limits were communicated to our data participants.

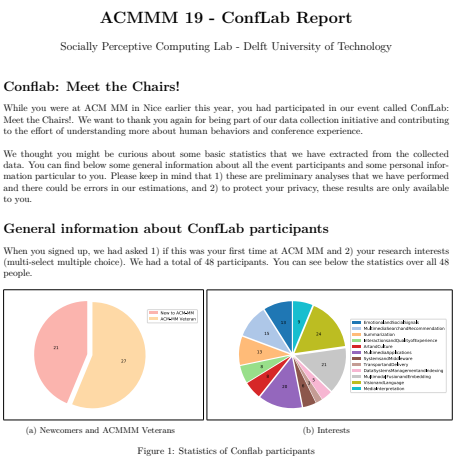
Q. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Only the latest version of the dataset will be maintained. If applicable, we will also host older versions of the data, accessible through the 4TU.ResearchData website.

Q. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We are open to contributions to the dataset. In accordance with our End User License Agreement, contributions should be made available, indicating if there are any restrictions on their contribution. We encourage the potential contributors to contact us to discuss how they wish to be attributed (e.g. citation of a paper or repository related to code/annotations). After finalizing the attribution discussion, we can add the attribution as an update following the same process explained above.

2.C Sample Participant Report



ACMMM 19 - ConfLab Report

Socially Perceptive Computing Lab - Delft University of Technology

Conflab: Meet the Chairs!

While you were at ACM MM in Nice earlier this year, you had participated in our event called ConfLab: Meet the Chairs!. We want to thank you again for being part of our data collection initiative and contributing to the effort of understanding more about human behaviors and conference experience.

We thought you might be curious about some basic statistics that we have extracted from the collected data. You can find below some general information about all the event participants and some personal information particular to you. Please keep in mind that 1) these are preliminary analyses that we have performed and there could be errors in our estimations, and 2) to protect your privacy, these results are only available to you.

General information about ConfLab participants

When you signed up, we had asked 1) if this was your first time at ACM MM and 2) your research interests (multi-select multiple choice). We had a total of 48 participants. You can see below the statistics over all 48 people.

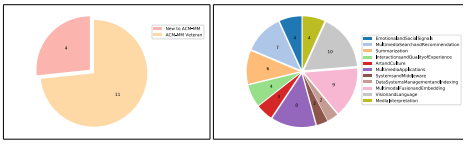
(a) Newcomers and ACMMM Veterans (b) Interests

Figure 1: Statistics of Conflab participants

1

Your networking behaviour - Bluetooth

Here we estimate how many people you have interacted with throughout the event. Our sensors record RSSI values and we set a single threshold for eliminating values corresponding to large physical distance that we do not consider as possible for face-to-face social interactions. We define the criterion of an interaction to be: 1) pairwise RSSI values below -55, and 2) pairwise proximity pings of at least 35 counted within a 1-minute window (sampling rate: 1Hz).



(a) Newcomers and ACMMM Veterans (b) Interests

Figure 2: Statistics of people you interacted with

In Figure 2a, the breakdown of the types of people you have interacted with is shown. In Figure 2b, you will find the interests breakdown of everyone you have interacted with. Figure 3 shows the distribution of the number of participants you interacted with. You will find yourself in the red bin; the x-axis says how many people you have interacted with and the y-axis says how many others had the same numbers as you.

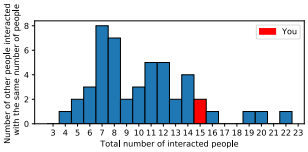


Figure 3: Distribution of the numbers of people participants interacted with

2

Your movement behavior - accelerometer

Here we estimate your motion behavior based on the accelerometer signal. Our sensors record tri-axial accelerometer values and we quantify the amount of motion by calculating the magnitude of the values of all 3 axes. We process the accelerometer data to separate movement and gravitational components of the signals based on a previous approach (Euclidean Norm Minus One [1]). For ease of visualization, we averaged the magnitude of acceleration over 30-second windows. You can see in Figure 4 your personal acceleration magnitude over time, as well as the mean and standard deviation values of acceleration magnitude for all participants over time.

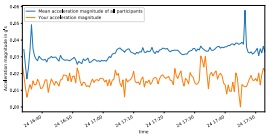
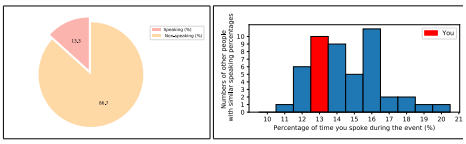


Figure 4: Acceleration magnitudes

Your speech behaviour - low-frequency audio

Here we estimate the amount of time you spoke. We first calculate the envelope of the low-frequency audio signal by taking the absolute value. Then, we apply a moving mean operator to the signal. By manually observing the signals of multiple participants, we selected a threshold to identify the speaking parts of the signal. We then further process the binary stream by filling the gaps between continuous speaking regions and eliminating speech regions that are smaller than a predefined threshold. Figure 5a and 5b show your percentage of speaking during the event and how you compare to the rest of the participants, respectively.

3



(a) Percentages of speech and non-speech (b) Distribution of speaking percentages for all participants

Figure 5: Your speaking behaviour

And that's it from the Socially Perceptive Computing Lab for now!

Note that for us, these analyses are just the starting point for estimating socially relevant behaviours. To do this more robustly and using more complex approaches is one of the reasons why we plan to share the data in next year or so. Maybe you are also curious to develop your own estimation techniques.

Finally, we welcome feedback on what other analyses that you are interested in, technical approaches, how to display your data better, your participatory experience, and any comments or advice that you might have for us. Please feel free to reply to this email or write to one of us directly.

Thanks again for your interest and we hope to see you again in the future!

[1] Bakrania, Kishan, et al. "Intensity thresholds on raw acceleration data: Euclidean norm minus one (ENMO) and mean amplitude deviation (MAD) approaches." PloS one 11.10 (2016): e0164045.

4

Figure 2.16: Sample post-hoc report sent to each participant of ConfLab. The report contains insights into the participant's networking behavior from the collected wearable-sensors data. This insight served as an additional incentive to participate in ConfLab, beyond interacting with the Chairs and contributing to a community-driven data endeavor (see main paper Section 2.3).

2.D Data Capture Setup Details

The Midge We improved upon the Rhythm Badge in three ways towards enabling more fine-grained and flexible data capture: (i) enabling full audio recording with a frequency up to 48 KHz, with an on-board switch to allow physical selection between high and low frequency capture directly at acquisition; (ii) adding a 9-axis Inertial Measurement Unit (IMU) with an on-board Digital Motion Processor (DMP) to record orientation; and (iii) an on-board SD card to directly store raw data, avoiding issues related to packet loss during wireless data transfer required by the Rhythm Badge. IMUs combine three tri-axial sensors: an accelerometer, a gyroscope, and a magnetometer. These measure acceleration, orientation, and angular rates respectively. These sensor measurements are combined on-chip by a Digital Motion Processor. Rough proximity estimation is performed by measuring the Received Signal Strength Indicator (RSSI) for Bluetooth packets broadcast every second (1 Hz) by every Midge. During the event, IMUs were set to record at 50 Hz. We recorded audio at 1250 Hz to mitigate extraction of verbal content while still ensuring robustness to cocktail-party noise.

Wireless Synchronization at Acquisition The central idea for our syncrhonization approach involves using a common Network Time Protocol (NTP) signal as reference for the camera and wearables sub-networks. The set-up achieved a cross-modal latency of 13 ms at worst, which is well below the 40 ms latency tolerance suitable for behavior research in our setting [18, Sec. 3.3]. Additionally, our synchronization approach allowed for dynamic addition of sensors to the network while still obtaining synchronized data streams. This is crucial in extreme in-the-wild events where some participants might arrive late.

Sensor Calibration For computing the camera extrinsics, we marked a grid of 1 m × 1 m squares in tape across the interaction area floor. We ensured line alignment and right angles using a laser level tool (STANLEY Cross90). For computing the camera intrinsics, we used the OpenCV asymmetric circles grid pattern [79]. The calibration was performed using the Idiap multi camera calibration suite [80]. All wearable sensors include one TDK InvenSense ICM-20948 IMU [81] unit that provides run time calibration. To establish a correspondence with the camera frame of reference, the sensors were lined up against a common reference-line visible in the cameras to acquire an alignment so that the camera data can offer drift and bias correction for the wearable sensors.

2.E Implementation Details

2.E.1 Person and Keypoint Detection Models

Data Cleaning A few frames contained some incorrectly labeled keypoints, a product of annotation errors like mis-assignment of participant IDs. We removed these using a threshold on the proximity to other keypoints of the same person. Further, in some cases, a person might be partially outside a camera's field of view. For the person detection task, we compute the bounding box from the keypoint ground-truth annotations. If more than half the body (50% keypoints) is missing in the frame so that e.g. only their legs are visible (see top of Figure 2.7a), we don't consider the person for that frame in the person detection experiments. Note that due to the significant overlap between the camera views, the person would be considered for the corresponding frame in the next camera. If they move back into the original view, we again take them into consideration for the original camera for the corresponding frame. Moreover, if there are more than 10% missing keypoints across all people in an image, we also discard that image from the experiment. This preprocessing resulted in a training set with 112k frames (1809k person instances) and a test set with 7k frames (158k person instances).

Training We resized the images to 960 × 540, and augmented the data by randomizing brightness and horizontal flips. The learning rate was set to 0.02 and batch size to 4. We trained the models for 50 k iterations, using the COCO-pretrained weights for initialization. All hyper-parameters were chosen based on the performance on a separate hold-out camera chosen as validation set. During training, any missing ground-truth keypoints (resulting from the person being partially outside the camera's view for instance) are ignored during back-propagation.

2.E.2 F-formation Detection

Data Cleaning Because keypoint annotations of the subjects are based on camera view and that the F-formation clustering methods cannot group subjects that do not exist under one camera view (e.g., when there are more identities than in associated ground truths), we processed the ground truth also based on camera number. This filtering pre-processing was decided based on the best camera view of the F-formations.

Feature Extraction The required features of GCFF and GTCG include location and orientation of the subjects. We used the X and Y position of subjects' head (as it is the most visible from the top-down view) for location, and extracted orientations for head, shoulders and hips. The orientations are calculated based on corresponding vectors determined by head and nose keypoints, left and right shoulder keypoints, and left and right hip keypoints, respectively.

Training We used pre-trained parameters for field of view (FoV) and frustum aperture (GTCG) and minimum description length (GCFF), provided in these models trained on the Cocktail Party. FOV and aperture are related to human eye gaze and head anatomical constraints reported by [82], and hence not dataset specific. The minimum description length is an initialized prior dictated by the same form of the Akaike Information Criterion, and becomes part of the optimization formulation. We tuned parameters such as frustum length (GTCG) and stride (GCFF) to account for average interpersonal distance in ConfLab based on Camera 6, as they vary across different datasets.

2.F Additional Results

2.F.1 Person and Keypoints Detection

Predictions from Pretrained SOTA Models Figure 2.17 shows predictions from SOTA human keypoint estimation models, namely, RSN [28], MSPN [83], HigherHRNet [84], and HourglassAENet [85], for the testing images of the Conflab dataset. Note that RSN and MSPN are top-down networks, i.e., they require person bounding boxes to predict the keypoints in each bounding box. We use COCO pretrained faster-RCNN network for bounding box estimation. HigherHRNet and HourglassAENet are bottom-up models, i.e., they directly predict keypoints from the full image. We use publicly available COCO pretrained checkpoints for prediction. The results show that the *state-of-the-arts 2D body keypoint detection models fail to capture the body keypoints in the Conflab dataset.* We infer that training on the dataset (e.g., COCO) that contains mostly side-view images does not work well in top-view images, for which Conflab dataset is important to the community.

Qualitative Results from ResNet-50 Finetuning Figure 2.18 illustrates more qualitative results from our finetuning experiments. We find that finetuning on our non-invasive top-down camera perspective significantly improves the keypoint estimation performance.

Ablations Tables 2.6 and 2.7 include the results of our experiments investigating the effect of varying the training data size on keypoint detection performance (see main paper Section 2.6.1). In Table 2.8, we show keypoint detection scores for experiments with different number of keypoints. We first focus on the five upper body keypoints: {head, nose, neck, rightShoulder, leftShoulder}. We then additionally considered the torso region keypoints for a total of nine: {rightElbow, rightWrist, leftElbow, leftWrist}. Finally, we add the hip keypoints {rightHip, leftHip} to the set. The experiments in the main paper are performed with all 17 keypoints. The results show that performance drops slightly when adding the arms keypoints ($5 \longrightarrow 9$, AP_{50}^{OKS} and AP^{OKS}), and that the relative gain when adding the hip keypoints ($9 \longrightarrow 11$) is lower than when adding the lower body keypoints ($11 \longrightarrow 17$, especially AP_{75}^{OKS}). We believe this is largely due to the lower body being more

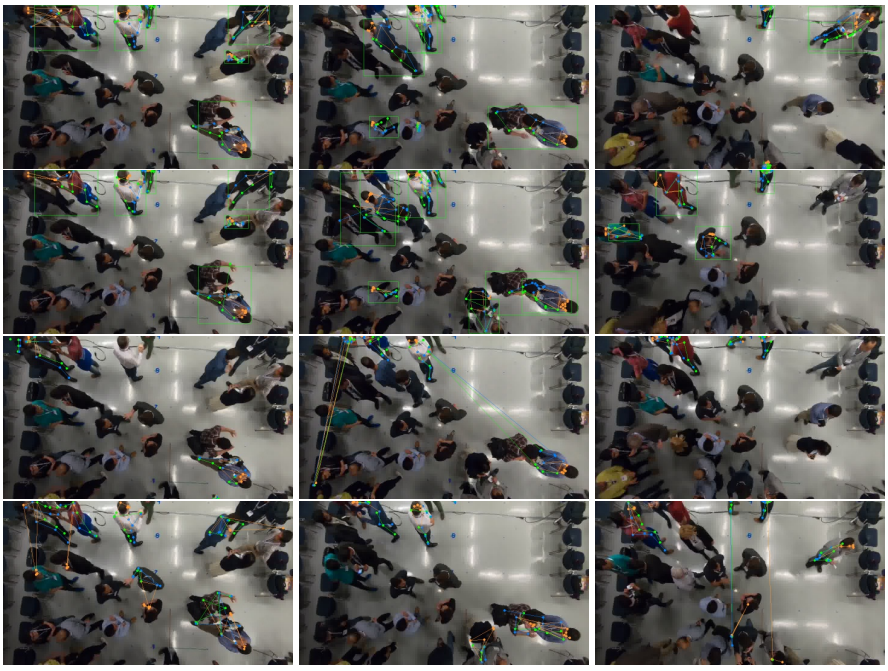


Figure 2.17: Results from Pretrained keypoint detection models. From top to bottom - predictions from RSN [28], MSPN [83], HigherHRNet [84], and HourglassAENet [85]. Results show that *SOTA 2D body keypoint detection models fail to capture the body keypoints in the ConfLab dataset.*

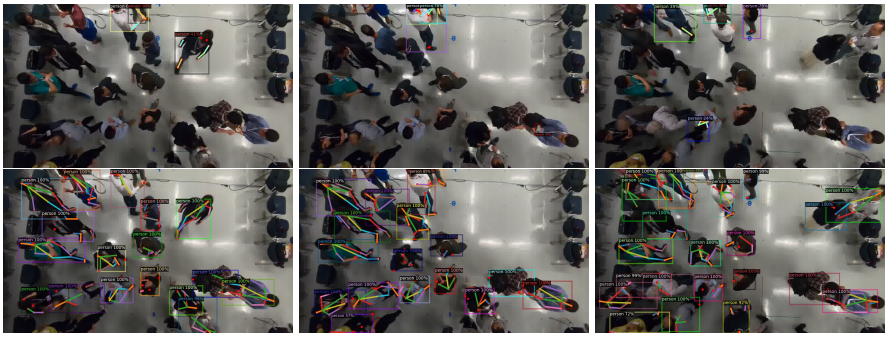


Figure 2.18: Results from (top) COCO pretrained Mask-RCNN model, (bottom) our ConfLab finetuned Mask-RCNN model.

static relative to the arms that move a lot to execute gestures during conversations.

2.F.2 Speaking Status Detection

Experiments with Different Sensor Modalities Table 2.9 displays the results from experiments using specific modalities from our IMUs for the task of speaking status detection.

Table 2.6: Effect of varying % frames from each camera at training on keypoint estimation.

% of training samples	AP^{OKS}_{50}
1.6%	29.0
3.2%	35.9
8%	39.0
16%	44.5
100%	45.3

Table 2.7: Effect of adding all frames from individual cameras to the training set on keypoint estimation.

Train Camera	#(training samples)	AP^{OKS}_{50}
cam 2	34k	8.6
cam 2 + cam 4	69k	31.1
cam 2 + cam 4 + cam 8	112k	45.3

Table 2.8: Keypoint estimation ablation with keypoints from different body sections: head and shoulders (5), + torso (9), + hips (11), + knees and feet (full 17).

#Keypoints	AP^{OKS}_{50}	AP^{OKS}	AP^{OKS}_{75}
5	26.6	7.1	1.4
9	26.5	6.9	2.0
11	35.8	9.5	2.2
17	45.3	13.5	3.3

Table 2.9: ROC AUC and accuracy for different sensor modalities from out 9-dof IMU in speaking status detection using the Minirocket classifier [64]. The number of channels in the corresponding modality is indicated in parentheses.

Input Modality	AUC	Accuracy
Acceleration (3)	0.813	0.768
Gyroscope (3)	0.765	0.716
Magnetometer (3)	0.610	0.656
Rotation vector (4)	0.726	0.696
All (13)	0.774	0.739

We used the best performing classifier (Minirocket [64]) among the ones tested in Table 2.3. The experiment setup is the same as detailed in Section 2.6.2, and the model is not changed between runs, except for the fact that different modalities may have a different number of input channels.

2.G Reproducibility Checklist

2.G.1 Person and Keypoints Detection

- Source code link: https://github.com/TUDelft-SPC-Lab/conflab
- Data used for training: 112k frames (1809k person instances).
- Pre-processing: See Section 2.4, Appendix 2.E.1.
- How samples were allocated for train/val/test: cameras 2, 4, and 8 are selected for training. For hyperparameter tuning, camera 8 are held out for validation.
- Hyperpatameter consideration: We considered learning rates (0.001/0.005/0.05/0.01), number of epochs (10/20/50/100), detection backbone (R50-FPN/R50-C4). Also see Appendix 2.E.1
- Number of evaluation runs: 5
- How experiments were ran: See Section 2.6.1.
- Evaluation metrics: Average precision at different thresholds.
- Results: See Section 2.6.1 and Appendix 2.F.1.

- Computing infrastructure used: All baseline experiments were ran on Nvidia V100 GPU (16GB) with IBM POWER9 Processor.

2.G.2 Speaking Status Detection

- Source code link: https://github.com/TUDelft-SPC-Lab/conflab
- Data used for training: 42884 windows (3 seconds), extracted from 48 participants' wearable data and speaking status annotations
- Pre-processing: Data was windowed into 3-second segments (see Section 2.6.2). The source code includes this pre-processing step.
- How samples were allocated for train/val/test: 10-fold cross-validation at the subject level (48 subjects) to test generalization to unseen data subjects. The splits can be reproduced exactly using the source code.
- Hyperparameter considerations: For acceleration-based methods, we used default network hyper-parameters and architectures from their tsai implementation [86]. For the MS-G3D baseline [61], we used default hyperparameters from the authors' implementation. For both, we determined the early stoppage point using a small subset (10%) of the training set.
- Number of evaluation runs: 1 run of 10-fold cross-validation
- How experiments were ran: For each fold, the early stoppage point was first determined using 10% of the training data as validation set and AUC as performance metric. The model at this stoppage point was then applied to the test set for evaluation.
- Evaluation metrics: Area under the ROC curve (AUC)
- Results: See Section 2.6.2
- Computing infrastructure used: Experiments were ran on a personal computer with GPU acceleration (NVidia RTX3080).

2.G.3 F-formation Detection

- Source code link: https://github.com/TUDelft-SPC-Lab/conflab
- Data used for training: Camera 6
- Pre-processing: See Section 2.E.2 for data cleaning and feature extraction.
- How samples were allocated for train/val/test: samples from Camera 6 were used to select the best model parameters. The rest are for test (evaluation). However, we note that Table 2.4 shows averaged performance on all cameras to provide a holistic view of the F-formation detection performance on ConfLab.
- (Hyper)parameter considerations: Both baseline methods are not deep-learning based and model parameters are interpretable. For GTCG, the parameters are frustum length (275), frustum aperture (160), frustum samples (2000), and sigma for affinity matrix (0.6). For GCFF, the parameters are minimum description length (30000) and

stride (70).

- Number of evaluation runs: 1
- How experiments were ran: A total of eight experiments were run for choosing the best parameters, and three for evaluation (for camera 2, 4, and 8). The parameters were chosen based on grid-search. For optimizing frustum length in GTCG, we searched over $[170, 195, 220, 245, 275]$ with 275 being averaged interpersonal distance based on Camera 6. For optimizing stride D in GCFF, we searched over $[30, 50, 70]$.
- Evaluation metrics: F1
- Results: See Section 2.6.3
- Computing infrastructure used: The experiments were run on Linux-based cluster instances on CPU with Matlab 2018a.

3

Synchronizing Multimodal Data at Acquisition for Capturing Social Interactions in the Wild

C. Raman*, S. Tan*, and H. Hung. A Modular Approach for Synchronized Wireless Multimodal Multisensor Data Acquisition in Highly Dynamic Social Settings. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, Seattle, WA, USA, 2020, pp. 3586-3594. DOI: 10.1145/3394171.3413697.

*Equal contribution

Abstract

Existing data acquisition literature for human behavior research provides wired solutions, mainly for controlled laboratory setups. In uncontrolled free-standing conversation settings, where participants are free to walk around, these solutions are unsuitable. While wireless solutions are employed in the broadcasting industry, they can be prohibitively expensive. In this work, we propose a modular and cost-effective wireless approach for synchronized multisensor data acquisition of social human behavior. Our core idea involves a cost-accuracy trade-off by using Network Time Protocol (NTP) as a source reference for all sensors. While commonly used as a reference in ubiquitous computing, NTP is widely considered to be insufficiently accurate as a reference for video applications, where Precision Time Protocol (PTP) or Global Positioning System (GPS) based references are preferred. We argue and show, however, that the latency introduced by using NTP as a source reference is adequate for human behavior research, and the subsequent cost and modularity benefits are a desirable trade-off for applications in this domain. We also describe one instantiation of the approach deployed in a real-world experiment to demonstrate the practicality of our setup in-the-wild.

Keywords: *synchronization, data collection, human behavior, social behavior, datasets*

3.1 Introduction

Human social behavior is a dynamic multimodal phenomenon; we express ourselves visually, vocally, and verbally. A significant focus of research here is the complex interpersonal dynamics between interaction partners, such as turn-taking in conversations [1, 2], or synchrony between participants [3]. An essential characteristic of these phenomena is their highly dynamic and multimodal nature; they evolve on short time-scales, requiring precise synchronization of multimodal and multisensor data streams.

Historically, human social behavior for automated analysis has been captured in controlled lab settings. As multimodal data analysis has become more prevalent, recorded sensors would be physically connected to relay timing information to ensure packet synchronization [4–6]. Concurrently, the ubiquitous computing community were developing approaches using wearable sensors that allowed for more pervasive sensing of social behaviors [7–9] while loosening strong requirements for data synchronization. As the trend moved towards more *in-the-wild* behavior analysis, multimedia researchers turned to collecting data in more uncontrolled settings that better matched real-world scenarios. Here, multiple visual and wearable sensing sources from both modalities have been combined [10, 11]. Figure 3.1 depicts a typical in-the-wild social interaction. In such prior works however, frame level synchronization requirements were circumvented by designing automated analysis approaches that smoothed behavioral data over broader time intervals on the order of a few seconds. On the other hand, the ubiquitous computing approach

has somewhat waived the need for more robust synchronization by adapting to problems that are able to take the wearable sensor data at face value and aggregate over sufficiently long time periods. This makes fine grained timing errors on the shorter scale of minutes or seconds less relevant [9].

In this paper, we argue that developing any approach to analyze the fine temporal dynamics of multi-modal multi-sensor behavioral data requires us to ensure a maximum temporal latency at the data collection stage of 40 ms (see Section 3.3.3 for further discussion). This requires us to bridge two traditions related to synchronization from the multimedia and ubiquitous computing domain which utilize different timing protocols and formats. Modalities such as audio and video, which have been used to analyze human behaviour analogous to human perception have used protocols such as PTP or GPS based reference time which enables sub-frame level synchronization using specialized hardware. Data here is often timestamped in the frame-based SMPTE timecode format such as linear time code (LTC)- HH:MM:SS:FF [12]. Meanwhile, in the ubiquitous computing domain, sensing devices have been born out of a tradition of wireless and distributed computing where each sensing device is itself also a microcomputer and as such has used NTP [13], relying on local UNIX system time to timestamp data. While it is widely understood that PTP or GPS based timing affords superior accuracy compared to NTP, setting up a multimodal multisensor system using the specialized hardware is prohibitively expensive.

In summary, we seek to answer the following question: how can we design a modular, cost-effective, distributed multi-sensor data acquisition setup for synchronized capture of social human behaviour in-the-wild? Concretely, our contributions are as follows:

- We propose and deploy a novel distributed data acquisition architecture built upon commercially available off-the-shelf components to wirelessly synchronize cameras (video) and wearable sensors (audio, inertial motion data, proximity) in-the-wild. Our core idea involves utilizing the Network Time Protocol (NTP) [14] as a common reference for all modalities, a choice contrary to conventional use in broadcasting setups.
- We show that the reduced accuracy of NTP in favor of significant cost and modularity benefits is a desirable trade-off for achieving crossmodal synchronization in data recording for human behavior research applications.

We support our argument in the rest of this work as follows. In Section 3.2 we review data recording or post-processing techniques used in other human behavior research and discuss the trade-offs involved. In Section 3.3 we establish acceptable latency tolerances for our application domain and propose our architecture, also describing a real-world instantiation of our system. We provide experiments to quantify the latency involved in our setup in Section 3.4 before discussing cost versus latency considerations in Section 3.5. Finally, we summarize our findings in Section 3.6.

Figure 3.1: A typical in-the-wild social interaction setting; adapted from the MatchNMingle Dataset [11]

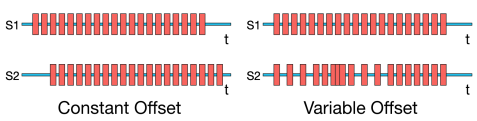


Figure 3.2: Basic types of desynchronization

3.2 Related Work

Synchronization Issues. We begin by first concretely describing the synchronization issues we propose to solve. We break these down into two basic types—constant and variable offset between data packets. Figure 3.2 depicts these issues for two data streams $S1$ and $S2$ over a world clock time axis t.

In the first case, all packets in $S2$ are offset from the corresponding packets in $S1$ by a uniform constant offset. This could arise because the triggers for recording the two streams are delayed, or because the internal clocks of the devices don't match. In the second case, while some packets are aligned in both streams, other packets are out of sync by a variable offset, and are said to have drifted. One such common scenario involves devices recording with variable framerate or dropped packets; for instance, while recording a long session with a standard webcam with autofocus or variable framerate, the video often drifts with respect to the audio over time. In practice, both these issues occur simultaneously, and information about the world clock is required to correct for these issues directly.

Event-based Approaches and Post Processing. Many widely used human behavior datasets attempt to fix the constant offset issues in post-processing by maximizing similarity scores around a manually identified common event in data streams. Traditionally, such an event included a balloon pop, a clap or the turning off of lights to get a common dark frame across cameras. More recently, Alameda-Pineda et al. use infra-red detections in cameras and wearable sensors to compute the optimal shift according to a similarity score [10]. Ringeval et al. use a common speech event such as the rise of a plosive to manually align high-quality audio from an external microphone to the low-quality audio from a webcam before computing the inter-correlation score around the located event [15]. While this approach helps with fixing mismatches around a single manually identified event, they are insufficient for fixing streams that have drifted over time or have variable offset (Alameda-Pineda et al. work with a no-drift assumption). More sophisticated approaches attempt to automatically identify events for synchronizing larger parts of the streams [16]. In contrast, we propose a modular approach that synchronizes the devices at data acquisition, requiring minimal—if any—post processing for synchronization.

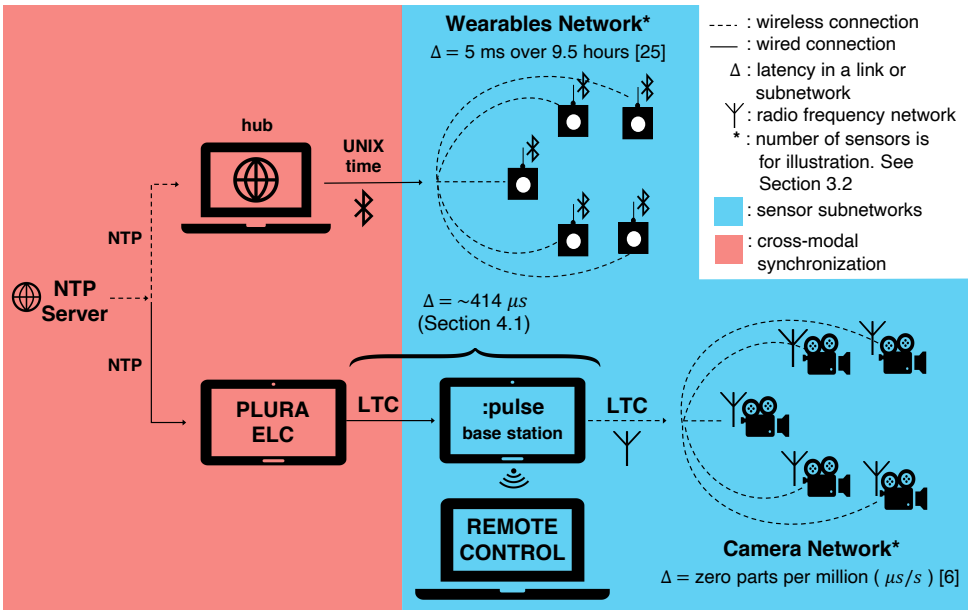


Figure 3.3: Overview of our proposed architecture. The reference time signal originates from the chosen NTP server and propagates to the subnetworks of wearable sensors and cameras.

Downstream Tasks. In addition to fixing synchronization issues in post-processing, a common approach is to mitigate their effect on downstream tasks. The core idea is to compute features over a window [17–20]. The size of this window is chosen to be larger than the duration by which the modalities are assumed to be out of synchronization. The features are computed using summary statistics, or by passing the individual features through a recurrent neural network and using the last hidden state as a representation of the window. This choice of window size, and whether this has a detrimental effect on the study of the phenomenon of interest can be contextualized by the discussion in Section 3.3.3.

Ubiquitous Computing Approaches. The analysis of social interactions has also been of interest to the ubiquitous computing community. Early work involved the development of custom wearable sensors like the UbER-Badge [21] to analyze interest and affiliation in conference attendees [22]. Period timestamps in these setups were relayed across a Radio Frequency (RF) network every 15 minutes. Cattuto et al. analyzed interactions in crowded social settings using custom RFID (Radio Frequency Identification) tags [23]. Packets from the tags were relayed to radio receivers that passed it to a central server for timestamping and storage. Their approach does not record timestamp at tag acquisition, and does not account for potential delays in transmission. For modeling longitudinal social interaction

networks in-the-wild, [8] used personal digital assistant (PDA) devices, and found the PDAs' clocks to be "shockingly unreliable", drifting up to 5 minutes across three weeks. Matic et al. infer interpersonal distance and relative orientation averaged over 10 s windows from up to five mobile phones in interactions lasting up to 15 minutes [20]. They state the mobile phones had synchronized clocks without specifying how they were synchronized.

Synchronization at Acquisition. A significantly more accurate, albeit expensive, approach compared to those discussed involves performing synchronization at data acquisition. This is achieved at the hardware level using either software or hardware triggers. Early approaches involved connecting low-cost cameras to standard computers over an Ethernet network and using software triggers to drive the recording [4, 5]. While the cost of sensors in these setups is low, the cost of computers remains. Timing control can be improved by using a common clock and physical hardware trigger lines into the cameras in an array [24], although this only works for the video modality.

Lichtenauer et al. significantly improved over previous works by proposing a system for multimodal data capture that centralizes the synchronization task by physically connecting the sensors to a multi-channel audio interface [6]. This approach was used in the recording of the MAHNOB-HCI datasets [25]. Other approaches have been proposed for setups involving motion-capture systems, where synchronization is achieved by plugging the output of the motion capture system to a robot in a human-robot interaction study [26], or in post-processing by performing an optimization over or manually annotated markers in a subset of frames [27]. These solutions are hard to deploy within in-the-wild settings over large physical areas since they are mainly wired solutions. They entail physically running trigger lines to the sensors of connecting the sensors or multiple PCs to a central audio interface. Comparatively, our solution affords for seamless decentralized addition of sensors to the system as long as those sensors are synchronizing clocks to the common NTP reference.

The closest work matching the scale and design requirements of our interaction setup is the MatchNMingle dataset [11], involving speed-dates followed by a mingling event. Their setup for the mingling event involves nine overhead GoPro cameras and wearable sensors on about 30 participants for each of three days. GoPro cameras in their setup are triggered using an infrared remote which might induce trigger delays, and no explicit timecode synchronization is done between the cameras which each record local time. The wearable sensors are synchronized intramodally to a global timestamp accurate to 1 second [28]. The video data is synchronized manually to the wearable sensors by using a GoPro to visually record the global timestamp propagating through the wearable network displayed on a screen. In contrast, our solution achieves timecode sync at acquisition at the microsecond level for the camera network and at the millisecond level across modalities.

To the best of our knowledge, the system we propose here is the first complete distributed and scalable multi-sensor data capture solution providing timecode synchronization between modalities at data acquisition for human behavior research.

3.3 Our Approach

Our core idea is to propagate a common time reference NTP signal to end devices (i.e., wearable sensors and cameras) at the time of data acquisition. Our approach is illustrated in Figure 3.3. The key challenge is that different subnetworks employ different timing information. The cameras use LTC for correct color framing and clock synchronization; the wearable sensors use the UNIX time received from the hub. With simply one additional hardware component (Plura ELC) combined with our choice of a common NTP reference, we achieve seamless crossmodal synchronization while preserving the existing local scheme of timekeeping. Starting from the origin of our system which is the NTP server, we explain the trade-offs of using NTP in Section 3.3.1. We describe a particular real-world instantiation of our system in Section 3.3.2, where we provide implementation details on how to relay time information to the sensor subnetworks. We contextualize latency measures within the human behavior research domain in Section 3.3.3, which frames our subsequent experimental design.

3.3.1 NTP as a Reference Signal

The main consideration of our approach is whether using NTP as a reference for cameras recording audiovisual data compromises the latency tolerance margins of the application when compared to more commonly used higher accuracy references such as PTP and GPS. Concretely, NTP is a software based protocol. While it uses a standardized, 64-bit UDP packet that can theoretically achieve picosecond timing, the latency error for NTP is heavily dependent on the network and ambient characteristics, and is typically measured on the order of milliseconds. On the other hand, PTP (specified in the IEEE 1588 standard) utilizes hardware based timestamping [29] to improve over NTP latency accuracy. With customized hardware, the latency error of PTP can be guaranteed to be on the order of microseconds. Though not as accurate as PTP or GPS-based solutions, using NTP has three advantages: firstly, *ease of setup*; synchronizing the system clock of a device to a local or public NTP server is straightforward, secondly, *modularity*; an entire subsystem of devices can be seamlessly added to the setup and guaranteed to be synchronized with all other devices if they synchronize to a common NTP reference, and thirdly, *reduced cost*; we discuss details in Section 3.5. For human behavior research applications, the lowered precision trade-off in favor of increased modularity of our setup is preferable, as we further contextualize in Section 3.3.3.

Specifically, the clock disciplining algorithm at the heart of the NTP specification states

that if left running continuously, an NTP client on a fast local area network in a home or office environment can maintain synchronization nominally within one millisecond [30]. As an implementation detail, practitioners can choose between a public server such as *time.google.com*, or an isolated local NTP server at the source. Using a local server avoids upstream latency introduced by network congestion. However, using a public server provides easier setup.

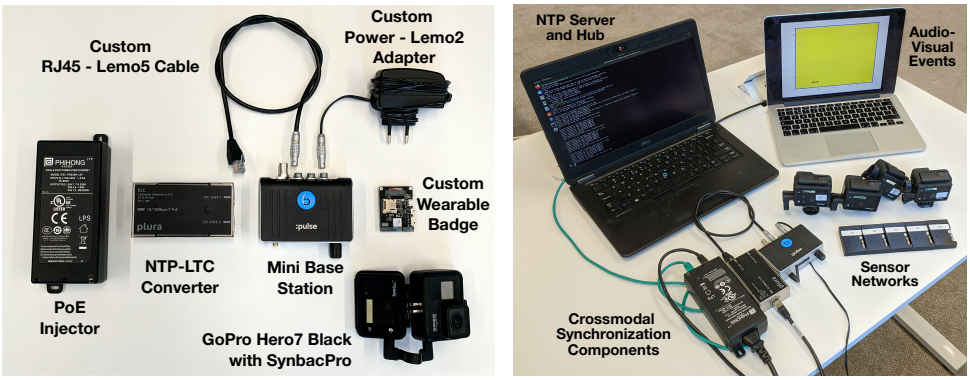
3.3.2 Real-World Implementation

We now describe one implementation of our approach. This setup was deployed to record data from a real-world social event. It involved 48 participants each wearing a sensor around their neck, in an interaction area of size 12m x 6m, captured by elevated and overhead cameras. Our setup included the following sensors:

- 13 GoPro Hero 7 Black video cameras (60fps, 1080p, Linear, NTSC) with audio (48 kHz); commercially available [31].
- 48 custom wearable sensors adapted from the open source Rhythm Badges [32]; each sensor includes an inertial measurement unit (IMU), mono microphone (1.2 kHz), and a Bluetooth proximity sensor.

The core components, custom hardware, and a working setup of our solution is depicted in Figure 3.4. Note that in keeping with privacy regulations, the wearable sensors record audio at frequencies only sufficient for detecting voice activity rather than verbal content. This makes the already subjective task of identifying semantic event boundaries in-the-wild even harder. Consequently, for the post-hoc evaluation of our system and comparison against widely used approaches in the domain that rely on such events for synchronization, we take a more principled approach to defining and sampling stimulus events, as we discuss in Section 3.4. While the number of devices we report here were used in our real-world deployment, it is not the system limit, as we discuss below. Our system is modular and scalable to larger number of devices with additional hubs and base stations (indicated in Figure 3.3).

Relaying Time to Cameras. We explain the bottom branch in Figure 3.3 regarding the camera network and its upstream components in this section. A laptop that receives the time reference from a local NTP server (same as the one used by the Bluetooth hub) shares the network time through a Power-Over-Ethernet injector (Plura 30W Single Port) with an Ethernet-to-LTC Converter (Plura ELC) [33]. The LTC signal that is converted from NTP is sent to a base station unit by Timecode Systems called :pulse [34], which allows for control, synchronization and metadata exchange for all devices within the camera network. It serves as the master in the localized master-slave radio frequency (RF) network, which shares its timecode with slave devices called Syncbac PRO [35], also manufactured by Timecode Systems. Each Syncbac PRO is physically tethered to a GoPro camera so



(a) Core components in our setup depicted only with custom cables; the connectors are aligned with the corresponding sockets.

(b) Full working setup of our data acquisition system, here shown with four cameras and five wearable sensors.

Figure 3.4: Real-world implementation of our proposed approach. Our working setup in Figure 3.4b is shown here recording audio-visual events for evaluating crossmodal synchronization, as discussed in Section 3.4.2.

that the accurate shared timecode is embedded within the MP4 files in each camera. In practice, once the timecode information of each video is available, any common video editing software can be used to align the video streams automatically for playback. An important consideration of our system design is to start the data acquisition remotely and wirelessly, since cameras are often mounted on the ceiling or other inaccessible places. The BLINK Hub app is used to remotely control (e.g. start, stop, etc), monitor and set features of all units within the localized RF network, which includes :pulse and Syncbac PRO. The BLINK Hub app can control up to 64 devices over a range of 500 m line of sight. Each :pulse unit can theoretically connect to an unlimited number of Syncbac PRO slaves within the same RF network over a range of 200 m line of sight. Both the RF network and the BLINK hub app control could have more network latency with increasing number of connections on the specific RF channel. The accuracy of the RF network synchronization is zero parts per million when the slaves (Syncbac PROs) are locked to the master (:pulse) [34, 35].

Note that our use of the ELC is different from its typical application of providing a signal for displaying the reference from a dedicated master reference generator. The novelty of our system stems from not requiring a typical GPS master reference generator at the source to phase lock to. Since our approach uses the local NTP server as the main reference itself, our use of the ELC allows for a simple method for video reference generation. Through experiments in Section 3.4 we show that our setup is appropriate for the domain. With the addition of a single component (any hardware or software NTP-LTC converter, the ELC in our setup), we wirelessly achieve crossmodal synchronization between the camera and wearables network compared to previous works as well as the more expensive GPS-based setup described in Section 3.4. Specifically, we are able to wirelessly embed the

timecode generated from the same reference used for other subnetworks into the video files, while relying on commercial products (with only custom connecting cables) for easier reproduction.

Relaying Time to Wearable Sensors. We explain the top branch in Figure 3.3 regarding the wearable sensors network in this section. Note that our system design is agnostic to the choice of the type of wearable sensors. Our choice of wearable sensors for this specific instantiation is motivated by the open source platform [32] for its accessibility and reproducibility, but could be replaced by any other subnetwork of sensors—wearable or otherwise—that supports NTP time synchronization. In our system, a hub node (in form of a laptop) receives the NTP time reference and shares it with the wearable sensors. The hub connects to the sensors sequentially in order of their MAC addresses for a Bluetooth handshake that transmits the UNIX time from the hub to the sensor. Each sensor then updates its system time to this timestamp. The frequency of establishing connection (i.e., synchronization messages) is a user defined parameter, and it has been shown that any interval between 0 and 600 seconds would be appropriate [36]. Since the hub is not maintaining a connection with all sensors at all times, there is no limit on the number of sensors that the hub can connect to. In practice, the maximum number of sensors associated to the hub is dictated by the saturation of wireless channel (i.e., when collisions occur). The mean average error in synchronization within the sensor network has been shown to be 5 ms over 9.5 hours of recording [36]. While intramodal synchronization within this subnetwork can be improved through various methods such as tracking the timestamps at each timestamp reception and parallelization of communication between the hub and the sensors, such improvements are outside the scope of our contribution.

We thereby achieve multisensor intramodal synchronization, multicamera intramodal synchronization, as well as multisensor-multicamera crossmodal synchronization. To summarize, each wearable is timestamped with the UNIX system time of the wearable network hub. The hub is set to the time of the local NTP server also providing time reference to the cameras, which are then recorded in terms of LTC. In post-processing, we convert the UNIX time to UTC time (HH:MM:SS:mS) to match samples to video frames denoted by LTC timecode (HH:MM:SS:FF). Note that these post-processing steps are insignificant compared to ones taken in manual alignment.

3.3.3 Latency Measures in Social Literature

To contextualize our assessment of tolerable latency margins, we review representative literature from social psychology that alludes to latency measures across different behavioral phenomena.

Measuring human response time (between stimulus and reaction) is an intuitive way to quantify behavior latencies. Early works have found that the response time spans

between 120 ms and 300 ms [37], with a specific example finding a 157 ms latency in speech perception [38]. Related to speech behavior is the more complicated turn-taking mechanism in conversations that involves pauses, gaps and overlaps. The time frame of consideration in identifying gaps between speakers. (speaker change) is approximately 200 ms, which is shown to be suitable for the task [1]. Studies in synchrony, mimicry, entrainment, and other higher-level social phenomena usually consider a larger window size. Levitan et al. have shown that a window size of 200-1000 ms works well in practice for studying speech backchannels. An episode of facial and body motor mimicry could be between 40 ms and 4 s [39, 40].

Apart from surveying the size of time frame used in various studies, an important measure of time offset is the latency in human perception of audiovisual data , since many human behavior datasets are manually annotated. Humans are shown to tolerate an audio lag of 200 ms or a video lag of 45 ms [41]. A successful automated method of data synchronization should perform on par with, if not better than human perception. It is worth noting that humans cannot annotate sensor data such as acceleration, in which case an automated synchronization solution is needed if aligning such data is required.

We deduce that offsets within a window size and/or range of human perception error, are generally tolerable. Based on the studies listed above, we consider a time offset to be acceptable if it is between 40 ms (e.g., facial analysis) and 1000 ms (e.g., entrainment). Though smaller offsets between different data streams can be achieved, the incremental gain becomes less relevant, especially for common phenomena of interest as discussed above. Nevertheless, our setup—in which we achieve a median video latency of 414 μs and wearable data latency of 5 ms over 9.5 hours [36]—is also applicable to data collection situations where fine details like faces are important such as egocentric vision setups, or those involving physiological sensors.

3.4 Experiments

The primary metric for synchronization accuracy is timing latency. A principled evaluation of our system would require characterizing latency at the local connection links in our proposed architecture, as well as final latency in the recorded data streams.

A common method for crossmodal synchronization used by human behavior datasets is the aligning of semantic events [10, 15]. As discussed in Section 3.3.2, given the subjective nature of start and end boundaries of semantic social events and low frequency audio recordings from wearables for privacy, we employ a more principled approach of defining and sampling stimulus ground-truth audio-visual events for our experiment presented in Section 3.4.2. Note that while the ground truth events are manually generated for control, the synchronization setup exactly matches the one we deployed in our in-the-wild

experiment.

Our core crossmodal approach introduces one point of latency through the use of an NTP-LTC converter to share the common NTP reference with the camera subnetwork. Since limited hardware connections prevent recording the output LTC streams during real-world deployment, we first present a pre-experiment to measure latency at the isolated connection in Section 3.4.1. Latency measures in our individual sensor subnetworks are depicted in Figure 3.3 and already discussed in Section 3.3.2.

With these time drifts quantified, we demonstrate that our approach is more robust and suitable for video, audio, and wearable sensor data alignment for the purpose of studying human behavior compared to previous approaches. Code and data for the decoding and analysis in these experiments are publicly available[1].

3.4.1 Timecode Latency between NTP-LTC Converter and Camera Network Master

We use the Plura Ethernet to LTC converter (ELC) for passing an LTC signal generated from the common NTP reference into the :pulse base station, as a timing reference for the camera network. In this experiment we evaluate the latency between two LTC signals: the LTC output of Plura ELC and the LTC output of :pulse.

Encoding. LTC is an encoding of timecode data within an audio signal. The timecode data is in the *hour:minute:second:frame* format. The data bits in an LTC signal are encoded using the biphase mark code (BMC) as depicted in Figure 3.5: a 0 bit has a single zero-one transition at the start of the bit period; a 1 bit has two transitions, at the beginning and middle of the bit period. Each LTC frame is made up of 80 bits of data, including a 16 bits long 'sync word' 0011111111111101 denoting the end of a frame. Consequently, at a framerate of 30 frames/sec, the LTC timecode has a maximum frequency of 2400 Hz (binary ones). In our experiments we measure the latency between the two LTC signals at the smallest possible time resolution; we consequently record the audio signals at the highest possible sampling frequency of 192 kHz, allowing for the smallest latency resolution of about 5 microseconds. Note that here theoretically, 80 audio samples correspond to 1 bit of data, and 80 bits correspond to 1 LTC frame.

Test Setup and Data. We passed the outputs of the Plura ELC (RJ45 jack) and the :pulse (BNC socket) to a Focusrite Scarlett 2i2 audio interface [42] through custom cables. Figure 3.7 depicts a part of our setup for recording the signals from the two devices. The Plura ELC was configured to use the public NTP server *time.google.com* as reference and generate an LTC signal at 30 frames/second. An isolated private NTP server can also be used upstream as mentioned, but that does not affect the outcome of the latency between

[1]Code & data are available at https://github.com/TUDelft-SPC-Lab/sync-experiments

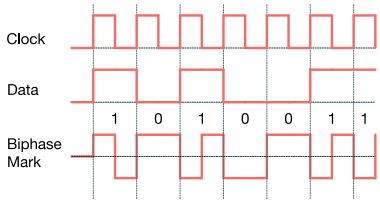


Figure 3.5: Biphase Mark Encoding of Linear Time Code

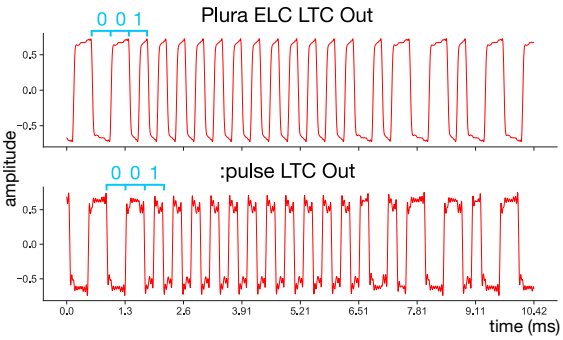


Figure 3.6: Raw audio LTC signals generated by the Plura ELC and :pulse modules. The window includes the encoding of an LTC sync word (0011111111111101) followed by the bits 0001000 from the next frame. The lower signal here leads the upper signal by 62 audio samples, or less than 1 bit of data.

the ELC and the :pulse we are studying here. The LTC signals were recorded using the application Audacity. We recorded for a total duration of 30 minutes over six sessions of five minutes each, for a total of 54000 LTC frames. Figure 3.6 depicts a window from our recorded audio signals at the end of a frame. The signals here represent the real-world noisy LTC signals encoded using the biphase mark code depicted in Figure 3.5.

Experiments. We measure synchronization at two levels: LTC frame level, and audio sample level. We use *demodulation* to refer to the conversion of the audio signal to binary data, and *decoding* to the conversion of the binarized data into the *hour:minute:second:frame* format. The recorded audio signals have imperfect leading and falling edges along with noise, with optima corresponding to a single data bit period being between 77-83 samples apart instead of the theoretical 80 audio samples. During demodulation, we begin by finding the local optima within a window size of six samples around the 80th sample following an optima. This new optima becomes the reference for the subsequent clock period. The demodulation was verified to match the original timecode presented in the recordings on the devices. We conducted a synchronization test using the 30 minutes of recording from six sessions where the binarized stream following the first sync word was decoded into timecode for checking correspondence at the frame level. We found that the data was indeed synchronized at the frame level for all the frames. With frame-level synchronization verified, we measured the world clock latency between the signals at the sub-frame level. We do this by finding the shift in number of audio samples to achieve maximum cross-correlation between the two audio signals. This lag was found to be [79, 80, 80, 80, -43, 78] samples for our six recordings, yielding a mean latency of 307.29 microseconds (59 samples) and a median latency of 414 microseconds (79.5 samples). A

positive lag implies that the :pulse signal leads the Plura ELC while a negative one implies the opposite. One way to interpret this is that the median latency is approximately 1 bit of data, which corresponds to 1/80th of an LTC frame. We conclude that this measure of latency is an order of magnitude lower than our overall acceptable latency tolerance of about 40 ms for the application domain as established in Section 3.3.3.

3.4.2 Evaluating Crossmodal Synchronization

Assuming that the GoPro audio and video are synchronized, we compare the audio recorded by the wearable sensors with the audio recorded by the GoPros in order to evaluate crossmodal synchronization of the wearable sensors and cameras of our system. We defined 10 stimulus audio-visual events that occurred randomly based on interval length (from 1-5 seconds) sampled from a Poisson distribution. An event is comprised of a visual color change accompanied by an audio *beep*. These events can be seen as the ground truth events in which the duration between each event is known. Figure 3.4b depicts our full working setup for recording these events.

The experiment considers 4 wearable sensor sensors and 4 GoPro cameras simultaneously capturing the generated audiovisual events played over approximately one minute. Figure 3.8 is a representative example showing that the audio events from one of the wearable sensors and one of the GoPro cameras appear to be in alignment. To further quantify the time offsets between different audio streams, we determine the number of samples between the end of an audio event and the onset of the subsequent event by thresholding the amplitude. Since the sampling frequencies of the wearable sensors (20 kHz) and the GoPros (48 kHz) are known, the number of samples is converted to time duration in seconds. We compare these empirically found durations from the recordings to ground truth durations between events .

We found that the average time offset for all wearable sensors and all GoPro recordings is 10.8±5.6 ms and 1.9±2.0 ms, respectively, when compared to the ground truth durations. Therefore, the maximum offset on average between wearable sensor and GoPro audio

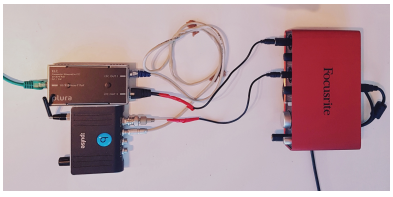


Figure 3.7: Hardware setup with custom cables for recording LTC signals from the Plura ELC and the :pulse base station.

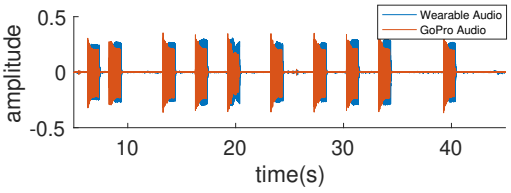


Figure 3.8: Representative example showing the aligned audio events in one of the wearable sensors and one of the GoPros.

signals is the sum of these offsets, resulting in approximately 13 ms, for a conservative estimation. In light of the latency in upstream links which are orders of magnitude smaller than what we observe here in the end devices, we offer some hypotheses on the possible sources of errors. Firstly, there is uncertainty in the generation and transmission of synchronization messages between the hub and the wearable sensors, ranging from a few milliseconds to several seconds, depending on connection interval settings [36, 43]. The time offset between the hub and the wearable sensors is inversely proportional to the frequency of connection. While it is possible to address this random time offset in Bluetooth connections via the Media Access Control (MAC) layer of the communication interface, the current approach is optimized towards energy efficiency [36]. Other possible reasons include varied quality of the wearable sensors and GoPro cameras resulting in discrepancy in sensor behavior and sensitivity, and offsets between the playback of the audiovisual events on the laptop (in Figure 3.4b) and the actual recording by the sensors. Despite the 13 ms offset across the camera and wearable sensor modalities, we highlight that it is still lower than both, the lower bound of 40 ms described in Section 3.3.3 and the human perception tolerance limit of audiovisual skew which is ± 80 ms [44]. In these purely perceptual tests, we could not hear any audible differences when the GoPro audio and the wearable sensor audio are played simultaneously. This shows that our approach is at least as good as, if not better than manual alignment of multimodal signals in the context of this experiment.

3.5 Cost versus Latency Considerations

Apart from providing a seamless interface for synchronizing different subnetworks of sensors, our choice of leveraging NTP as the common reference is also motivated by cost—the only component we have introduced to achieve crossmodal synchronization is the NTP to LTC converter. We have also shown that the reduced accuracy of our choice is well within tolerable latencies between sensors for our application domain. But what if cost is not a constraint?

For setups enjoying higher budgets, we recommend using synchronization references from highly-accurate GPS satellites. These satellites are all synchronized to the same time using stabilized atomic clock hardware and known locations due to their medium earth orbits. As a result, GPS receivers can listen to multiple broadcast sources and use trilateration (somewhat similar to triangulation) to determine their own position and time deviation. GPS modules can consequently perform time-synchronization with a resolution of 100 nanoseconds or smaller [45].

Through the use of satellites, a GPS based solution largely mitigates issues like unquantifiable delays in network communications or a lack of local operating system resources

commonly plaguing the use of the protocols described in Section 3.3.1. Additionally, GPS modules can be used to generate NTP and PTP signals [46] for downstream subnetworks. One potential downside of using GPS references is that the GPS antenna needs to be installed outdoors under visible sky to obtain the GPS reference, which might pose logistical challenges depending on the physical setting of the interactions being studied.

Since we use the Plura ELC in our setup, for comparison we provide an example GPS controlled setup using components from Plura. This involves modules from their Rubidium Series [47]. A GPS receiver such as the RUB G16X would obtain the GPS signal and pass it as reference to the RUB GT master timecode generator module to produce an LTC signal. This LTC signal would act as an external reference for the :pulse base station like in our current setup. A RUB PM-N module connected to the the GT would serve the dual purpose of powering the setup and acting as an NTP server to generate the NTP signal for the hub of the wearable sensor network similar to our current setup. The entire setup would be housed in a RUB H1 rack. The GPS setup for crossmodal synchronization is approximately eight times more expensive than our setup using an ELC and a POE injector[2].

3.6 Conclusion

In this paper we introduce a novel approach for synchronized and wireless acquisition of human behavior data across video, audio, and wearable sensor data modalities, captured in highly dynamic in-the-wild settings. The key challenge of synchronization in these settings is to propagate a common time reference signal to end devices such as cameras and wearable sensors in a wireless and scalable manner without compounding network delays. Another challenge is that different types of sensors rely on different types of timing information. Existing solutions in this space are either wired solutions, or achieve limited synchronization in post-processing, making them less suitable for our scenario involving a large number of people free to move in a large physical area. Our novel solution uses a common NTP reference signal for both the camera and wearable sensors modalities; conventionally NTP is superceded by more accurate reference signals for video. Through empirical experiments, we show that the median time latency introduced by our choice of using NTP is 414 μs for the video modality. The intramodal latency of our wearable sensor network built by extending an open platform is 5 ms over 9.5 hours [36]. The overall crossmodal latency of our setup is approximately 13 ms at worst based on an events-based experiment. We contextualized our findings using latency measures from representative social behaviour literature, and find that our setup performs well within a tolerable latency margin of 40 ms for our application domain and human perception.

[2]The GPS setup described currently costs approximately US $5700, while the combined cost of the ELC and the POE injector is about US $730.

To the best of our knowledge, this is the first work that quantifies latency tolerances for a data collection system designed for collecting human behavior data, and proposes a distributed architecture built on commercially available products. Through valid trade-offs, our approach provides a practical, accurate, cost-effective, time-efficient, and modular solution that is more advantageous than the current state-of-the art methods/heuristics for highly dynamic social settings.

Acknowledgments

This research was partially funded by the Netherlands Organization for Scientific Research (NWO) under the MINGLE project number 639.022.606. We thank Ruud de Jong, Jeroen Bastemeijer, Amelia Villegas, Paul Scurrell, Thomas Rock, Jürgen Loh, and Ekin Gedik for sharing their technical expertise and giving us helpful feedback.

References

[1] M. Heldner and J. Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38 (4):555–568, 2010.

[2] R. Levitan, A. Gravano, and J. Hirschberg. Entrainment in speech preceding backchannels. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, pages 113–117. Association for Computational Linguistics, 2011.

[3] E. Delaherche, M. Chetouani, A. Mahdhaoui, et al. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012.

[4] X. Cao, Y. Liu, and Q. Dai. A flexible client-driven 3dtv system for real-time acquisition, transmission, and display of dynamic scenes. *EURASIP Journal on Advances in Signal Processing*, 2009:1–15, 2008.

[5] T. Svoboda, H. Hug, and L. Van Gool. Viroom—low cost synchronized multicamera system and its self-calibration. In *Joint Pattern Recognition Symposium*, pages 515–522. Springer, 2002.

[6] J. Lichtenauer, J. Shen, M. Valstar, and M. Pantic. Cost-effective solution to synchronised audio-visual data capture using multiple sensors. *Image and Vision Computing*, 29(10):666–680, 2011.

[7] T. Choudhury and A. Pentland. Sensing and modeling human networks using the sociometer. In *Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings.*, pages 216–222. IEEE, 2003.

[8] D. M. Wyatt et al. *Measuring and modeling networks of human social behavior.* University of Washington, 2010.

[9] D. O. Olguín, B. N. Waber, T. Kim, et al. Sensible organizations: Technology and methodology

for automatically measuring organizational behavior. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):43–55, 2008.

[10] X. Alameda-Pineda, J. Staiano, R. Subramanian, et al. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8): 1707–1720, 2015.

[11] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung. The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, 2018.

[12] St 12-1:2014 - smpte standard - time and control code. *ST 12-1:2014*, pages 1–41, 2014.

[13] D. Mills, J. Martin, J. Burbank, and W. Kasch. Network time protocol version 4: Protocol and algorithms specification. 2010.

[14] D. L. Mills. Internet time synchronization: the network time protocol. *IEEE Transactions on communications*, 39(10):1482–1493, 1991.

[15] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013.

[16] D. Bannach, O. Amft, and P. Lukowicz. Automatic event-based synchronization of multimodal data streams from wearable and ambient sensors. In *European Conference on Smart Sensing and Context*, pages 135–148. Springer, 2009.

[17] E. Gedik and H. Hung. Detecting conversing groups using social dynamics from wearable acceleration: Group size awareness. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–24, 2018.

[18] C. Raman and H. Hung. Towards automatic estimation of conversation floors within f-formations. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 175–181. IEEE, 2019.

[19] A. Rosatelli, E. Gedik, and H. Hung. Detecting f-formations & roles in crowded social scenes with wearables: Combining proxemics & dynamics using lstms. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 147–153. IEEE, 2019.

[20] A. Matic, V. Osmani, and O. Mayora-Ibarra. Analysis of social interactions through mobile phones. *Mobile Networks and Applications*, 17(6):808–819, 2012.

[21] M. Laibowitz, J. Gips, R. AyIward, A. Pentland, and J. A. Paradiso. A sensor network for social dynamics. In *2006 5th International Conference on Information Processing in Sensor Networks*, pages 483–491. IEEE, 2006.

[22] J. Gips and A. Pentland. Mapping human networks. In *Fourth Annual IEEE International Conference on Pervasive Computing and Communications (PERCOM'06)*, pages 10–pp. IEEE, 2006.

[23] C. Cattuto, W. Van den Broeck, A. Barrat, et al. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one*, 5(7), 2010.

[24] S. Tan, M. Zhang, W. Wang, and W. Xu. Aha: An easily extendible high-resolution camera array. In *Second Workshop on Digital Media and its Application in Museum Heritages (DMAMH 2007)*, pages 319–323, 2007.

[25] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.

[26] E.-S. Fotinea, E. Efthimiou, A.-L. Dimou, et al. Data acquisition towards defining a multimodal interaction model for human–assistive robot communication. In *International Conference on Universal Access in Human-Computer Interaction*, pages 613–624. Springer, 2014.

[27] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010.

[28] M. C. Dobson. *Low-power epidemic communication in wireless ad hoc networks.* PhD thesis, Vrije Universiteit, 2013.

[29] J. Eidson and K. Lee. Ieee 1588 standard for a precision clock synchronization protocol for networked measurement and control systems. In *Sensors for Industry Conference, 2002. 2nd ISA/IEEE*, volume 10. Ieee, 2002.

[30] D. Mills. Clock discipline algorithm, 2014. URL `https://www.eecis.udel.edu/~mills/ntp/html/discipline.html`.

[31] Go pro hero 7 black. `https://gopro.com/en/nl/shop/cameras/hero7-black/CHDHX-701-master.html`.

[32] O. Lederman, D. Calacci, A. MacMullen, et al. Open badges: A low-cost toolkit for measuring team communication and dynamics. *arXiv preprint arXiv:1710.01842*, 2017.

[33] Plura ethernet to ltc convertor. `https://plurainc.com/wp-content/uploads/2019/03/eELCmanual.pdf`.

[34] Timecode systems mini-basestation. URL `https://www.timecodesystems.com/wp-content/uploads/2016/08/Pulse-manual-Web-1.1-1.pdf`.

[35] Timecode systems syncbacpro. `https://www.timecodesystems.com/syncbac-pro/`.

[36] M. Hopfengaertner. An open-source sensor platform for analysis of group dynamics. *arXiv preprint arXiv:1901.04977*, 2018.

[37] R. D. Luce et al. *Response times: Their role in inferring elementary mental organization.* Number 8. Oxford University Press on Demand, 1986.

[38] C. A. Fowler, J. M. Brown, L. Sabadini, and J. Weihing. Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of memory and language*, 49(3):396–413, 2003.

[39] M. Sonnby-Borgström, P. Jönsson, and O. Svensson. Emotional empathy as related to mimicry

reactions at different levels of information processing. *Journal of Nonverbal behavior*, 27(1): 3–23, 2003.

[40] S. Bilakhia, S. Petridis, A. Nijholt, and M. Pantic. The mahnob mimicry database: A database of naturalistic human interactions. *Pattern recognition letters*, 66:52–61, 2015.

[41] K. W. Grant, V. v. Wassenhove, and D. Poeppel. Discrimination of auditory-visual synchrony. In *AVSP 2003-International Conference on Audio-Visual Speech Processing*, 2003.

[42] Focusrite scarlett-2i2. URL https://focusrite.com/en/usb-audio-interface/scarlett/scarlett-2i2.

[43] Bluetooth core specifications. https://www.bluetooth.com/specifications/bluetooth-core-specification/.

[44] R. Steinmetz. Human perception of jitter and media synchronization. *IEEE Journal on selected Areas in Communications*, 14(1):61–72, 1996.

[45] E. Sazonov, V. Krishnamurthy, and R. Schilling. Wireless intelligent sensor and actuator network-a scalable platform for time-synchronous applications of structural health monitoring. *Structural Health Monitoring*, 9(5):465–476, 2010.

[46] P. Volgyesi, A. Dubey, T. Krentz, et al. Time synchronization services for low-cost fog computing applications. In *2017 International Symposium on Rapid System Prototyping (RSP)*, pages 57–63. IEEE, 2017.

[47] Plura inc. rubidium series. URL https://www.plurainc.com/solutions/timers/rubidium-series/.

4

Synthesizing Training Data for Face-Related Tasks

C. Raman, C. Hewitt, E. Wood, and T. Baltrušaitis. Mesh-Tension Driven Expression-Based Wrinkles for Synthetic Faces. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, Hawaii, USA, 2023, pp. 3515-3525.

® T. Baltrusaitis, **C. Raman**, C. Hewitt, and E. Wood. Face image generation with wrinkles. (Patent Application.)

Abstract

Recent advances in synthesizing realistic faces have shown that synthetic training data can replace real data for various face-related computer vision tasks. A question arises: how important is realism? Is the pursuit of photorealism excessive? In this work, we show otherwise. We boost the realism of our synthetic faces by introducing dynamic skin wrinkles in response to facial expressions and observe significant performance improvements in downstream computer vision tasks. Previous approaches for producing such wrinkles either required prohibitive artist effort to scale across identities and expressions, or were not capable of reconstructing high-frequency skin details with sufficient fidelity. Our key contribution is an approach that produces realistic wrinkles across a large and diverse population of digital humans. Concretely, we formalize the concept of mesh tension and use it to aggregate possible wrinkles from high-quality expression scans into albedo and displacement texture maps. At synthesis, we use these maps to produce wrinkles even for expressions not represented in the source scans. Additionally, to provide a more nuanced indicator of model performance under deformations resulting from compressed expressions, we introduce the 300W-winks evaluation subset and the Pexels dataset of closed eyes and winks.

4.1 Introduction

Synthetic data has been commonly employed for a variety of computer vision tasks including object recognition [2–5], scene understanding [6–9], eye tracking [10, 11], hand tracking [12, 13], and full body analysis [14–16]. However, the complexity of modeling the human head has largely precluded the generation of full-face synthetics for face-related machine learning. While realistic digital humans have been created for movies and video games, they usually entail significant artist effort per character [17, 18]. Consequently in literature, the synthesis of facial training data has been accompanied by simplifications, or a focus on parts of the face such as the eye region [19, 20] or the *hockey mask* [21–24]. This has resulted in a *domain gap*—a difference in distributions between real and synthetic facial data that makes generalization challenging. Efforts towards bridging this domain gap have mainly utilized domain adaptation to refine synthesized images [25] or domain-adversarial training where models are encouraged to ignore domain differences [26]. As such, generating realistic face data has been considered so challenging that it is assumed that synthetic data cannot fully replace real data for in-the-wild tasks [25].

To directly address the challenge, Wood et al. [1] attempted to minimize the domain gap at the source, by generating synthetic faces with unprecedented realism. Their method procedurally combines a parametric 3D face model with a comprehensive library of high-quality artist-created assets including textures, hair, and clothing. In doing so, the method overcomes a key bottleneck in techniques employed by the Visual Effects (VFX) industry for

Figure 4.1: Final renders for a diverse set of synthetic identities and expressions. For each identity we illustrate renders using the base method of Wood et al. [1] (left), and our added technique for generating expression-based wrinkling effects (right). For the same expression parameters, our method produces varied wrinkling effects across distinct identities (middle and bottom row).

synthesizing realistic humans—that of scale. The procedural sampling can randomly create and render novel 3D faces without manual intervention. Machine learning systems trained on the synthesized data for landmark localization and face parsing achieved performance comparable with the state-of-the-art without using a single real image.

However, one limitation of the method proposed by Wood et al. [1] is the lack of dynamic, expression dependent wrinkles. The method generates textures using only the neutral-expression scans, which remain static for all deformations of the underlying face mesh resulting from expression changes. In this work we propose a simple yet effective method for incorporating expression-based wrinkles. Our central idea is to capture complex wrinkling effects for an identity from high-resolution scans of their posed expressions. We store all these possible wrinkles into albedo and displacement textures we refer to as *wrinkle maps*. At synthesis, for any arbitrary expression beyond those represented in the source scans, we blend between the neutral and wrinkle textures using a notion of the *tension* in the face mesh to obtain dynamic wrinkling effects. Figure 4.1 contrasts the results of our method against the current state-of-the-art (SOTA) approach for face synthetics. We also include animated sequences in the Supplementary Material.

The term *wrinkle maps* was first used by early VFX approaches to refer to artist-defined bump or normal maps for simulating animated wrinkles [27–30]. However, these approaches suffer from three drawbacks. First, the bump and normal maps only *simulate* underlying geometry changes; the silhouette and shadows which are of relevance for face related tasks such as landmark localization remain unaffected. Second, the methods do not affect the albedo or diffuse textures. Finally, the most crucial drawback is scale. The methods entail manual definition of wrinkle maps and masks for every blendshape for every character. In contrast, our automatic mesh-tension driven method naturally scales with the number of identities and expressions, while incorporating real wrinkles for both albedo and displacement textures from scans. Furthermore, we also handle identities without expression scans, transferring plausible wrinkles from the most similar neutral textures.

To advance the development of synthetics for face-related tasks, we make the following concrete contributions:

- A system for dynamic, expression-based wrinkles that scales easily with increasing identities and expressions.
- A demonstration of empirical qualitative and quantitative improvement over the SOTA synthetics system on face-keypoint localization and surface-normal estimation.
- Novel evaluation data and metrics for keypoint localization in the eye region where wrinkles are especially relevant for learning tasks.

4.2 Background: Synthesizing Faces

We build upon the work of Wood et al. [1] for synthesizing face images for downstream machine learning tasks. Their method involved sampling from a generative 3D blendshape-based face model learned from 3D scans of 511 individuals with neutral expression. The sampled face is then *dressed up* with samples from a large collection of hair, clothing, and accessory assets. For each synthesized face, the authors employ three textures that remain fixed across all expressions: one albedo map for skin color; one coarse displacement map to encode scan geometry not captured by the sparsity of the vertex-level identity model; and one meso-displacement map to approximate skin-pore level detail built by high-pass filtering the albedo texture. In contrast, we automatically compute an additional sets of albedo and displacement wrinkle textures from expression scans to support dynamic wrinkling effects.

4.3 Related Work

Wrinkle Maps Oat [27] proposed using a pair of bump maps to render animated wrinkles on virtual characters. These bump maps—called *wrinkle maps*—store surface normals for an

expanded (or stretched) and compressed (or *scrunched-up*) expression, typically obtained from artist sculpted high-resolution meshes. A base normal map stores fine surface details such as pores. In order to achieve independently controlled wrinkles, the face is divided into multiple regions. Each region is specified by an artist-defined mask stored in a texture map. An animated scalar wrinkle weight in the range $[-1, 1]$ then interpolates between the two wrinkle maps for each masked region: at either end of the range one of the wrinkle maps is at its full influence, with a weight of 0 corresponding to no influence on the base normal map. A similar method was later independently proposed by Duque Reis et al. [31] using a single wrinkle map. Jimenez et al. [29] expanded on the scheme proposed by Oat [27], allowing for the use of any number of wrinkle maps, with a weight in the range of $[0, 1]$ defining the influence of each map. Subsequent improvements to make the technique amenable in real-time or performance driven settings involved the dynamic generation of either the region masks [30] or the wrinkle weights [28]. Both approaches relied on using a *skinned* mesh attached to bones. Dutreve et al. [30] proposed generating dynamic region masks by using the bone influence weights from a set of artist defined reference poses. Oat [28] proposed generating dynamic wrinkle weights by comparing each mesh triangle's area before and after skinning, a technique derived from Microsoft's DirectX 10 Sparse Morph Targets demo [32]. While the term *wrinkle maps* in literature has been alternatively used to refer to bump or normal maps, in this work we use the term to collectively refer to the textures used for synthesizing wrinkles: the albedo and displacement maps corresponding to the expanded and compressed textures.

Simulation Based Approaches While the use of wrinkle maps is the most common methodology when artistic control is of importance, several alternate techniques have been proposed for simulating wrinkles on 3D surfaces. These methods can broadly be grouped into physical and geometric simulation of wrinkles. An early physical simulation based approach employed a biomechanical perspective, considering the skin as an elastic membrane and modeling the deformations using linear plastic model [33]. Boissieux et al. [34] extended the elastic membrane perspective by modeling the skin as a volumetric substance comprising layers of different materials and using a finite element method for computing deformations. Finite element modeling was also employed in subsequent works to simulate forearm skin wrinkling [35], and skin aging [36]. Wang et al. [37] and Venkataraman et al. [38] proposed energy based approaches. Here, wrinkle deformations are produced by minimizing an energy function indicating flexure properties of a governing curve on a surface. To produce wrinkles on dynamic meshes such as simulated cloth, Müller and Chentanez [39] proposed attaching a higher resolution wrinkle mesh to the coarse base mesh and determining the deviations of the wrinkle mesh vertices using a static solver [40]. Geometric simulation based approaches typically involve expressing the wrinkles using some geometric primitives. Bando et al. [41] represented wrinkles using a cubic Bezier

curve, generating their furrows from a sequence of starting points along a user specified direction field. Other proposed techniques involved the use of length preserving constraints on planar curves along with artist placed features at locations on an animated mesh where wrinkling is desired [42, 43]. Ilie et al. [44] employed a Hermite spline interpolation along with a modified Rayleigh distribution function to simulate wrinkling activity in facial animations. Subsequent methods extracted wrinkle curves automatically from images [45, 46]. Finally, Gui et al. [47] used both a muscle model and a geometric wrinkle shape function to simulate 3D facial wrinkles.

Machine Learning Approaches More recently, several methods for expression and texture synthesis, and facial performance capture have addressed the synthesis of wrinkles. As part of their performance capture system, Cao et al. [48] trained regressors for mapping local image appearance to wrinkle displacements to augment a coarse face mesh tracked in real-time. Zeng et al. [24] and Richardson et al. [22] proposed convolutional networks based refinement architectures to reconstruct detailed facial geometry from a single image. Nagano et al. [49] proposed a conditional generative adversarial network architecture for the synthesis of image-based dynamic 3D avatars. Given a single neutral-face input image, their system can generate novel photo-real expressions from alternate viewpoints, including variable details such as wrinkles. More directly, Deng et al. [50] proposed a variational autoencoder architecture to synthesize plausible fine-scale wrinkles on a variety of coarse-scale 3D faces.

4.4 Synthethizing Expression-Based Wrinkles

Figure 4.2 illustrates an overview of our approach. The underlying idea is that wrinkles can by synthesized additively over the neutral-expression textures. We formalize the concept of mesh tension and use it to automatically aggregate wrinkling effects in a data-driven manner across all expression scans of an identity. We store these possible wrinkles corresponding to the expansion and compression deformations of the face in separate albedo and displacement textures, which we collectively refer to as wrinkle maps in this work. Note that displacement maps modify the underlying geometry unlike bump or normal maps that simply simulate the geometry changes. At synthesis, we sample a face mesh from a generative face model [1] and randomly select a set of neutral and wrinkle textures corresponding to an identity from the available scans. We then compute the tension in the face mesh to drive the blending between the neutral and wrinkle maps to obtain dynamic wrinkling effects. In contrast with previous learning-based wrinkling methods [22–24, 50], we do not build a generative model for the textures since such models struggle to reconstruct high frequency details such as wrinkles compared to directly extracting them from scans.

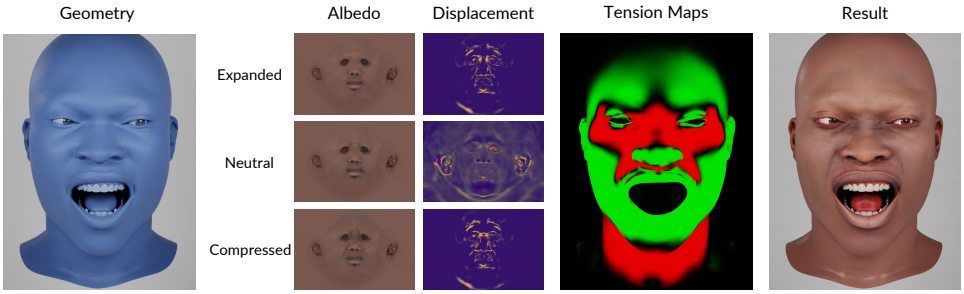


Figure 4.2: Method Overview. The state-of-the-art method for face synthetics [1] generates albedo and displacement textures using only the neutral-expression scan for an identity (middle row, also see Figure 4.1). In contrast, we automatically compute expanded and compressed texture maps to aggregate wrinkling effects in the face and neck regions across available posed-expression scans for the identity. At synthesis, for a given set of arbitrary expression parameters we compute the local tension at every vertex in the corresponding face mesh: we depict expansion in green and compression in red. This mesh tension serves as weights to dynamically blend between the neutral, expanded, and compressed texture maps to synthesize the wrinkling effect at that vertex. Note that our method can thereby generate wrinkles for expressions even beyond those represented in the source scans.

4.4.1 Mesh Tension

We formalize mesh tension to capture the amount of compression or expansion at each vertex of a 3D polygon mesh resulting from a deformation. More concretely, we express mesh tension as a function of the mean change in the length of the edges connected to a vertex as a result of the deformation. Consider an undeformed mesh $\overline{\mathbf{X}} = (\overline{V}, \overline{E})$ with a sequence of vertices $\overline{V}$ and sequence of edges $\overline{E}$, that undergoes a deformation to result in the mesh $\mathbf{X} = (V, E)$. We only consider deformations such that $\overline{\mathbf{X}}$ and $\mathbf{X}$ possess the same topology. For vertex $v_i \in V$, let $(e_1, \ldots, e_K)$ denote the sequence of K edges connected to v_i, with $(\overline{e}_1, \ldots, \overline{e}_K)$ denoting the corresponding edges in $\overline{X}$ connected to $\overline{v}_i$. We then define the mesh tension at v_i as

$$t_{v_i} \coloneqq 1 - \frac{1}{K} \sum_{k \in [K]} \frac{\|e_k\|}{\|\overline{e}_k\|}, \tag{4.1}$$

where $[K] = \{1, \ldots, K\}$, and $\|.\|$ denotes edge length. Note that we subtract from 1 so that positive values of t_{v_i} indicate compression, negative values indicate expansion, and a value of 0 indicates no change.

In practice, for finer manual control we introduce the parameters of strength s to scale the tension, and bias b to artificially favor expansion or compression, computing the weighted tension at v_i as $t'_{v_i} = s \cdot t_{v_i} + b$. Further, we allow for artificial propagation of expansion and compression effects through the mesh. For each effect we introduce a parameter denoting the number of iterations for a morphological dilation (positive values) or erosion (negative values) operation. The propagation of each effect is first performed independently

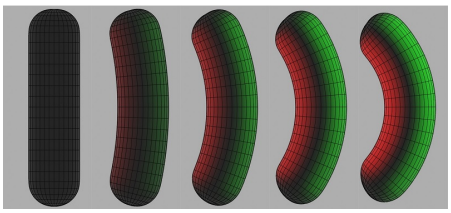


Figure 4.3: Mesh Tension. We illustrate our computation of mesh tension for various deformations of a simple cylinder. Expansion is depicted in green and compression in red. Black shading corresponds to zero tension.

Figure 4.4: Data—High-resolution 3D Scans. For each identity, we illustrate: the raw neutral scan (top-left), the manually-cleaned neutral scan to remove sensor noise and hair (top-right), and two raw expression scans (bottom).

over the mesh, and the resulting tension values are added for vertices that end up with both expansion and compression. Figure 4.3 illustrates these effects for a simple cylindrical mesh. See Appendix 4.A for additional illustrations of the effect of the tension parameters. Code as a Blender [51] add-on is available at https://github.com/chiragraman/mesh-tension.

4.4.2 Data and Preprocessing

We start with a set of high-quality commercially available 3D scans of 208 individuals. All 208 identities contain scans with neutral expressions, while 52 contain additional scans for posed expressions. The neutral scans were manually cleaned for removing noise and hair artifacts, and registered to the topology of the 3D face model proposed by Wood et al. [1], resulting in a mesh of $7,667$ vertices and $7,414$ polygons. Figure 4.4 illustrates the scans.

Automatic Cleaning of Expression Scans The manual cleaning of scans is a labor-intensive process. To automate the process of masking the noise and hair artifacts from the expression scans, we utilize the difference between the raw and manually-cleaned neutral scans. Concretely, we employ a two-stage masking procedure illustrated in Figure 4.5. First, we apply an identity-agnostic coarse mask to filter most artifacts outside of the hockey-mask and neck regions where expression-based wrinkling occurs. Next, to capture the manual changes made by the artists in the cleaning of each neutral scan, we employ a Gaussian Mixture Model-based background subtraction technique [52]. Treating the clean neutral textures as background and the raw original ones as foreground, we obtain an identity-specific mask of the noise and hair artifacts for every identity. We apply this fine mask to clean the textures from the corresponding expression scans for each identity.

4.4.3 Data-Driven Wrinkle Maps

Tension-Weighted Wrinkle Maps Figure 4.6 illustrates our method for generating wrinkle maps from the face scans. Our underlying idea is to use the tension at each vertex

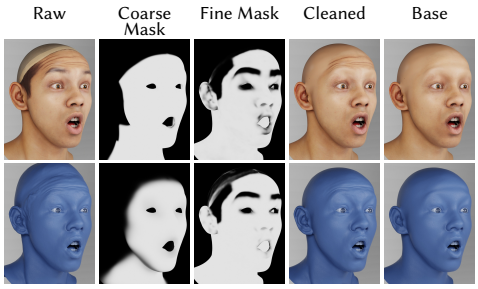


Figure 4.5: Cleaning Raw Textures. We illustrate the cleanup of albedo (top) and displacement (bottom) textures on the *surprise* expression. We automatically remove the hair and sensor noise artifacts in the raw textures around the head, neck, and cheeks while preserving the desired wrinkles in the nose, forehead, and mouth regions (compared to the base mesh, with neutral albedo and without displacement respectively, for the same expression).

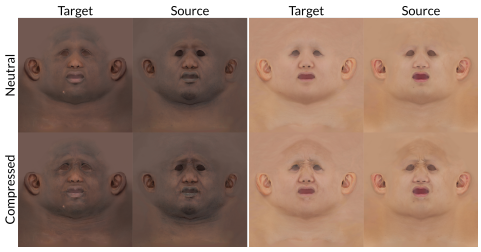


Figure 4.7: Grafting Wrinkles. For an identity with missing expression scans (target), we find the identity from among those with expression scans that has the most similar neutral albedo map (source). We then graft the wrinkles from the source's wrinkle map onto the target's neutral texture to obtain the target wrinkle maps (here illustrating the compressed albedo).

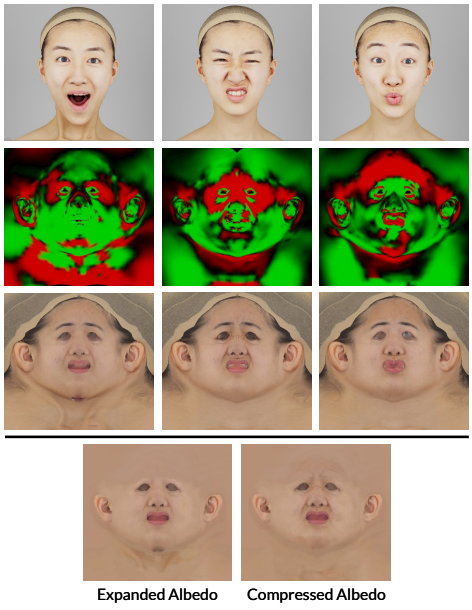


Figure 4.6: Generating Wrinkle Maps from Scans. We illustrate the computation of albedo wrinkle maps with three raw expression scans (top). We compute the tension maps corresponding to the scans (middle), depicting expansion in green and compression in red. Finally, the expression albedo textures (bottom) are linearly combined using the normalized tension as weights to obtain the expanded and compressed albedo wrinkle maps. A similar procedure is applied to obtain the displacement wrinkle maps.

Figure 4.8: Final Renders for Some Identities with Grafted Wrinkles. We computed the wrinkle maps for these identities by grafting wrinkles from identities with expression scans (see Figure 4.7). We illustrate two expressions for each identity, without (left) and with wrinkles (right).

as weights in a linear combination of the cleaned textures across expressions, with zero tension corresponding to the neutral textures. (Figure 4.6 depicts raw textures for easier visual correspondence with the scans.) We begin by fitting the generative face model from Wood et al. [1] to the raw scans and compute the tension maps from the resulting meshes. The individual expansion and compression maps are then normalized using the softmax function. Finally, we linearly combine expression textures using the normalized tension as weights to obtain the expanded and compressed wrinkle maps. The same procedure is applied to obtain both the albedo and displacement wrinkle maps.

Identities With Missing Expression Scans How do we compute wrinkle maps for the identities without posed expression scans? We employ a simple wrinkle-grafting procedure. For a target identity without wrinkle maps, we find the source identity with wrinkle maps that has the most similar neutral albedo map, measured by mean squared error in pixel color. For the source identity, we compute the wrinkling effects as the difference between the neutral and wrinkle maps (for both albedo and displacement). We then add this difference to the neutral textures for the target identity to obtain the target wrinkle maps. We illustrate the grafting procedure for the compressed albedo maps in Figure 4.7, and final example renders with grafted wrinkles in Figure 4.8.

4.5 Experiments and Results

We evaluate our proposed mesh-tension driven wrinkles both quantitatively and qualitatively on two face analysis tasks: landmark detection (Section 4.5.1) and normal estimation (Section 4.5.2). We compare adding mesh-tension to the existing SOTA method for full-face synthetics, and compare the performance of models trained on the resulting data against SOTA approaches in the field for these tasks.

4.5.1 Landmark Localization

Experimental Details. We use direct regression based facial landmark detection [53] with an off-the-shelf ResNet 101 [54]. We use a 256 × 256 px RGB image as input to predict 703 dense facial landmarks. We additionally employ label translation [1] to deal with systematic inconsistencies between our 703 predicted dense landmarks and the 68 sparse landmarks labeled as ground truth in our evaluation datasets (this is done only for Table 4.1).

As a training dataset we rendered $100k$ synthetic images, consisting of $20k$ identities with 5 frames for each identity (different view-points, expressions, and environments). We also generated ground-truth annotations of 703 dense 2D landmarks from the face-meshes to accompany each image. We train our models for 300 epochs using PyTorch Lightning, starting with a learning rate of 1e−3 and halved every 100 epochs.

Table 4.1: Landmark Localization on 300W. We normalize mean error using interocular distance. Lower is better.

Method	Common NME	Challenging NME	Private $FR_{10\%}$
Trained on Real Data			
LAB [55]	2.98	5.19	0.83
AWING [56]	**2.72**	**4.52**	0.33
ODN [57]	3.56	6.67	-
3FabRec [58]	3.36	5.74	**0.17**
LUVLi [59]	2.76	5.16	-
Trained on Synthetic Data			
No wrinkles [1]	3.11	4.84	0.33
Ours (wrinkles)	3.10	4.83	**0.17**

Table 4.2: Landmark Localization - Eyes. We report eye-opening errors for Pexels, and eyelid point-to-polyline errors for 300W and the *winks* subset. In all cases normalized by bounding-box diagonal. Lower is better.

Method	Pexels	300W	300W-winks
Trained on Real Data			
AWING [56]	1.06	0.62	**0.69**
3FabRec [58]	3.60	0.81	1.32
Trained on Synthetic Data			
No wrinkles [1]	0.97	0.51	0.86
Ours (wrinkles)	**0.86**	**0.48**	0.74

Evaluation Datasets and Metrics. We use the **300W** dataset [60] (with common, challenging and private subsets), and employ the standard normalized mean error (NME) and failure rate ($FR_{10\%}$) error metrics [60].

While the 300W dataset provides evaluation of overall landmark detection performance, it is not sensitive enough to detect improvements in specific parts of the face or during particular expressions. We identify a small subset of 30 images from 300W that contain winks and compressed face expressions (**300W-winks**) to provide a more nuanced indication of performance under such deformations. We report errors for eyelid-landmarks by taking a point-to-line distance from every predicted eyelid landmark to the corresponding polyline defining an eyelid in ground truth. This metric allows us to better understand eye region error and to use different landmark definitions in training and evaluating models (e.g. from our 703 landmark model or from 98 landmark models [56]). See Appendix 4.E for the list of images in 300W-winks.

We also introduce a **Pexels** dataset which contains 318 images of fully closed eyes (because of blinking, scrunching or compressing the face) and 105 images with only a single eye closed (winking). This allows us to asses model performance under such conditions which are rare in other datasets. To collect the data we used a stock photography website [1] using search terms *wink/blink/compress/scrunched* and similar image searches. We select only semi-frontal images with no or limited occlusion of the eyes to best evaluate performance in that region. The URLs of the images selected can be found in Appendix 4.F. Knowing which images contain fully closed eyes or just a single eye closed allows us to measure eyelid accuracy without explicit landmark annotations. We define the eye opening error as the mean eye aperture of both eyes in the *eye-closed* case and eye aperture of closed eye in the *wink* case. See Appendix 4.B for illustrations of the above two metrics.

[1] https://www.pexels.com/

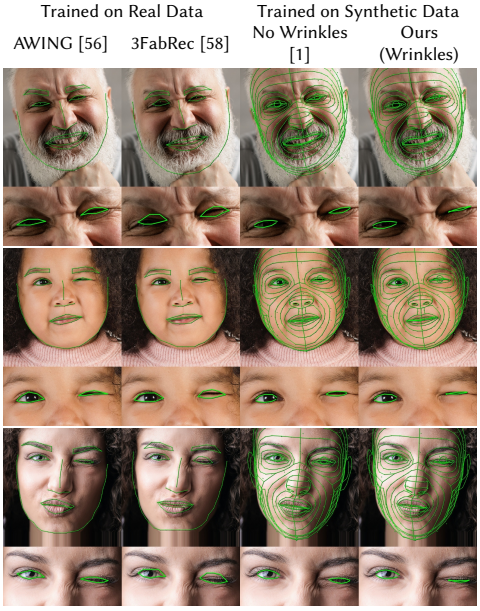


Figure 4.9: Qualitative results for landmark localization on Pexels. Training on synthetic faces with our expression-based wrinkles is crucial for localizing keypoints in compressed regions of the face.

Figure 4.10: Expression-Based Wrinkle Components. We add wrinkles through two components: displacement and albedo. Here we show each in isolation. Displacement is critical for achieving realistic lighting of wrinkles. Especially note the forehead (zoomed) and neck regions.

Table 4.3: Landmark Localization Ablation. We report eye-opening errors for Pexels, and eyelid point-to-polyline errors for 300W and the *winks* subset. Lower is better.

Dataset	Base	Disp. Only	Albedo Only	Full
300W	0.51	0.51	0.50	**0.48**
300W-winks	0.86	0.76	0.80	**0.74**
Pexels	0.97	**0.86**	0.89	**0.86**

Baselines. We compare against recent SOTA methods trained on images of real faces. For subsequent nuanced analysis on 300W-winks and Pexels we consider the methods of Wang et al. [56] and Browatzki and Wallraven [58] since they collectively yield the best performance on 300W.

Results. From Table 4.1 we see that our proposed mesh-tension driven wrinkles provide a marginal improvement for landmark localization. However, when we look at specific eye region results on 300W, 300W-winks and Pexels in Table 4.2, we see that improvement is much larger for the eye region and our synthetic-only trained approaches outperform real-data based models. Also see Figure 4.9 and Appendix 4.C.

Ablation. We further analyze the importance of the albedo and displacement wrinkling components for landmark detection. From Figure 4.10 and Table 4.3 we see that displacement plays a more important role than albedo in improving performance, but best results are achieved through a combination of both.

4.5.2 Surface-Normals Prediction

Surface normals can be used to infer 3D information about a surface from 2D images, and have been used in several human-centered vision tasks such as clothing [61] and

Figure 4.11: Qualitative Surface-Normals Predictions on Pexels. The model trained on synthetic faces with wrinkles recovers significantly more high-frequency details.

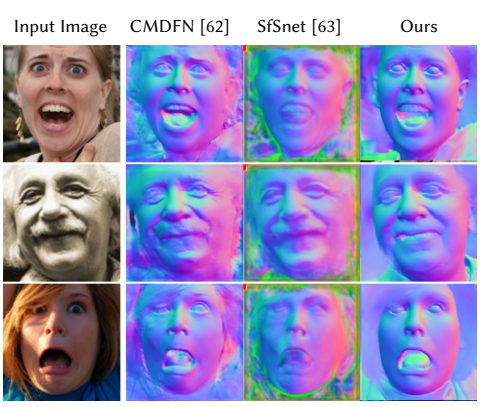


Figure 4.12: Qualitative Comparison against SOTA. Our synthetic data-only U-Net yields predictions comparable to SOTA while being less noisy and more robust to lighting.

face-shape [62] reconstruction and relighting [63].

We train a U-Net [64] with a ResNet 18 [54] encoder to predict camera-space surface normals of the face. As input we use 256×256 px RGB images from a dataset of $50k$ synthetics images. The network is trained for 200 epochs using cosine similarity loss with a learning rate of 1e−3. Camera-space surface normal images rendered as part of our synthetic data pipeline are used as ground-truth.

Results on real images are shown in Figure 4.11; the network trained on images synthesized with our method recovers more high-frequency detail on the face. As shown in Figure 4.12, we achieve comparable results to other recent methods for face surface-normals prediction [62, 63]. Further comparisons are provided in Appendix 4.D.

4.6 Conclusion

We have presented a method for introducing dynamic expression-based wrinkles to synthetic faces that yields improved performance on the downstream tasks of landmark localization and surface-normals estimation, especially for regions of the face most deformed by expressions.

Our use of tension in the face mesh is key in the automatic scaling of our method with identities and expressions, which has been a bottleneck for past wrinkling approaches that rely on prohibitive artist effort. In addition, our data-driven approach also enables the capturing of real wrinkles from scans which doesn't require artistic judgment.

By boosting the realism of synthesized faces with dynamic wrinkles, we have made an explicit case for synthetic data: our method yields improved performance for models

on downstream tasks. In addition, synthesizing data with diverse faces across races and genders involves significantly less effort than collecting representative datasets in the wild. Consequently, downstream real-life systems developed using such synthetic data are less likely to suffer from unfair biases along these sensitive variables.

Acknowledgments

Chirag would like to thank: Tom Cashman, Stephan Garbin, and Panagiotis Giannakopoulos for the insightful discussions; Sebastian Dziadzio for help with fitting the face model; Sarah Roberts for being an infallible remover of obstacles; and Steve Miller (BA: @shteeve) for an initial implementation of the Blender mesh tension add on.

References

[1] E. Wood, T. Baltrušaitis, C. Hewitt, et al. Fake It Till You Make It: Face analysis in the wild using synthetic data alone. *arXiv:2109.15102 [cs]*, Oct. 2021.

[2] Y. Yao, L. Zheng, X. Yang, M. Naphade, and T. Gedeon. Simulating Content Consistent Vehicle Datasets with Attribute Descent. *arXiv:1912.08855 [cs]*, July 2020.

[3] T. Hodan, V. Vineet, R. Gal, et al. Photorealistic Image Synthesis for Object Instance Detection. In *arXiv:1902.03334 [Cs]*, pages 66–70, Feb. 2019.

[4] W. Qiu, F. Zhong, Y. Zhang, et al. UnrealCV: Virtual Worlds for Computer Vision. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, pages 1221–1224, New York, NY, USA, Oct. 2017. Association for Computing Machinery. doi: 10.1145/3123266.3129396.

[5] A. Rozantsev, V. Lepetit, and P. Fua. On Rendering Synthetic Images for Training an Object Detector. *Computer Vision and Image Understanding*, 137:24–37, Aug. 2015. doi: 10.1016/j.cviu.2014.12.006.

[6] A. Kar, A. Prakash, M.-Y. Liu, et al. Meta-Sim: Learning to Generate Synthetic Datasets. *arXiv:1904.11621 [cs]*, Apr. 2019.

[7] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. *arXiv:1605.06457 [cs, stat]*, May 2016.

[8] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for Data: Ground Truth from Computer Games. *arXiv:1608.02192 [cs]*, Aug. 2016.

[9] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, Las Vegas, NV, USA, June 2016. IEEE. doi: 10.1109/CVPR.2016.352.

[10] E. Wood, T. Baltrusaitis, X. Zhang, et al. Rendering of Eyes for Eye-Shape Registration and Gaze Estimation. *arXiv:1505.05916 [cs]*, May 2015.

[11] L. Świrski and N. Dodgson. Rendering synthetic ground truth images for eye tracker evaluation. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 219–222, Safety Harbor Florida, Mar. 2014. ACM. doi: 10.1145/2578153.2578188.

[12] F. Mueller, F. Bernard, O. Sotnychenko, et al. GANerated Hands for Real-time 3D Hand Tracking from Monocular RGB. *arXiv:1712.01057 [cs]*, Dec. 2017.

[13] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. *arXiv:1704.07809 [cs]*, Apr. 2017.

[14] G. Varol, J. Romero, X. Martin, et al. Learning from Synthetic Humans. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, July 2017. doi: 10.1109/CVPR.2017.492.

[15] J. Shotton, A. Fitzgibbon, M. Cook, et al. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pages 1297–1304, June 2011. doi: 10.1109/CVPR.2011.5995316.

[16] Huazhong Ning, Wei Xu, Yihong Gong, and T. Huang. Discriminative learning of visual words for 3D human pose estimation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, AK, USA, June 2008. IEEE. doi: 10.1109/CVPR.2008.4587534.

[17] D. Hendler, L. Moser, R. Battulwar, et al. Avengers: Capturing thanos's complex face. In *ACM SIGGRAPH 2018 Talks*, pages 1–2, Vancouver British Columbia Canada, Aug. 2018. ACM. doi: 10.1145/3214745.3214766.

[18] Karis, Brian, Antoniades, Tameem, Caulkin, Steve, and Mastilovic, Vladimir. Digital Humans: Crossing the Uncanny Valley in UE4. In *Game Developers Conference*, 2016.

[19] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 131–138, Charleston South Carolina, Mar. 2016. ACM. doi: 10.1145/2857491.2857492.

[20] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828, Columbus, OH, USA, June 2014. IEEE. doi: 10.1109/CVPR.2014.235.

[21] M. Sela, E. Richardson, and R. Kimmel. Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation. *arXiv:1703.10131 [cs]*, Mar. 2017.

[22] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning Detailed Face Reconstruction from a Single Image. *arXiv:1611.05053 [cs]*, Apr. 2017.

[23] E. Richardson, M. Sela, and R. Kimmel. 3D Face Reconstruction by Learning from Synthetic Data. *arXiv:1609.04387 [cs]*, Sept. 2016.

[24] X. Zeng, X. Peng, and Y. Qiao. DF2Net: A Dense-Fine-Finer Network for Detailed 3D Face Reconstruction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2315–2324, Seoul, Korea (South), Oct. 2019. IEEE. doi: 10.1109/ICCV.2019.00240.

[25] A. Shrivastava, T. Pfister, O. Tuzel, et al. Learning from Simulated and Unsupervised Images

through Adversarial Training. *arXiv:1612.07828 [cs]*, July 2017.

[26] Y. Ganin, E. Ustinova, H. Ajakan, et al. Domain-Adversarial Training of Neural Networks. *arXiv:1505.07818 [cs, stat]*, May 2016.

[27] C. Oat. Animated wrinkle maps. In *ACM SIGGRAPH 2007 Courses*, SIGGRAPH '07, pages 33–37, New York, NY, USA, 2007. Association for Computing Machinery. doi: 10.1145/1281500.1281667.

[28] C. Oat. Real-Time Wrinkles, 2007.

[29] J. Jimenez, J. I. Echevarria, C. Oat, and D. Gutierrez. Practical and Realistic Facial Wrinkles Animation. In *GPU pro 2*, chapter Practical and Realistic Facial Wrinkles Animation. AK Peters Ltd., 2011.

[30] L. Dutreve, A. Meyer, and S. Bouakaz. Real-Time Dynamic Wrinkles of Face for Animated Skinned Mesh. In *ISVC' 09: 5th International Symposium on Visual Computing*, Advances in Visual Computing, pages 25–34, Las Vegas, USA, United States, Nov. 2009. Springer. doi: 10.1007/978-3-642-10520-3_3.

[31] C. Duque Reis, G. Reis, J. De Martino, and H. Batagelo. *Real-Time Simulation of Wrinkles*. Feb. 2008.

[32] M. D. . S. Team. Sparse Morph Targets Sample, 2007.

[33] Y. Wu, P. Kalra, and N. M. Thalmann. Physically-based Wrinkle Simulation & Skin Rendering. In *Computer Animation and Simulation '97*, pages 69–79. Springer Vienna, Vienna, 1997. doi: 10.1007/978-3-7091-6874-5_5.

[34] L. Boissieux, G. Kiss, N. M. Thalmann, and P. Kalra. Simulation of Skin Aging and Wrinkles with Cosmetics Insight. In *Computer Animation and Simulation 2000*, pages 15–27. Springer Vienna, Vienna, 2000. doi: 10.1007/978-3-7091-6344-3_2.

[35] C. Flynn and B. A. O. McCormack. Finite element modelling of forearm skin wrinkling. *Skin Research and Technology*, 14, 2008.

[36] C. Flynn and B. A. O. McCormack. Simulating the wrinkling and aging of skin with a multi-layer finite element model. *Journal of Biomechanics*, 43(3):442–448, Feb. 2010. doi: 10.1016/j.jbiomech.2009.10.007.

[37] Y. Wang, C. C. Wang, and M. M. Yuen. Fast energy-based surface wrinkle modeling. *Computers & Graphics*, 30(1):111–125, 2006.

[38] K. Venkataraman, S. Lodha, and R. Raghavan. A kinematic-variational model for animating skin with wrinkles. *Computers & Graphics*, 29(5):756–770, Oct. 2005. doi: 10.1016/j.cag.2005.08.024.

[39] M. Müller and N. Chentanez. Wrinkle meshes. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 85–92, 2010.

[40] M. Müller, B. Heidelberger, M. Hennix, and J. Ratcliff. Position based dynamics. *J. Vis. Comun. Image Represent.*, 18(2):109–118, Apr. 2007. doi: 10.1016/j.jvcir.2007.01.005.

[41] Y. Bando, T. Kuratate, and T. Nishita. A simple method for modeling wrinkles on human skin.

In *10th Pacific Conference on Computer Graphics and Applications, 2002. Proceedings.*, pages 166–175, Beijing, China, 2002. IEEE Comput. Soc. doi: 10.1109/PCCGA.2002.1167852.

[42] C. Larboulette and M.-p. Cani. Real-Time Dynamic Wrinkles, 2004.

[43] M. Li, B. Yin, D. Kong, and X. Luo. Modeling Expressive Wrinkles of Face For Animation. In *Fourth International Conference on Image and Graphics (ICIG 2007)*, pages 874–879, Aug. 2007. doi: 10.1109/ICIG.2007.22.

[44] M. D. Ilie, C. Negrescu, and D. Stanomir. A robust mathematical model for simulating wrinkle activity in 3D facial animations. In *2012 10th International Symposium on Electronics and Telecommunications*, pages 271–274, Nov. 2012. doi: 10.1109/ISETC.2012.6408082.

[45] L. Li, F. Liu, C. Li, and G. Chen. Realistic wrinkle generation for 3D face modeling based on automatically extracted curves and improved shape control functions. *Computers & Graphics*, 35(1):175–184, Feb. 2011. doi: 10.1016/j.cag.2010.08.003.

[46] R. Vanderfeesten and J. Bikker. Example-Based Skin Wrinkle Displacement Maps. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 212–219, Parana, Oct. 2018. IEEE. doi: 10.1109/SIBGRAPI.2018.00034.

[47] J. Gui, Y. Zhang, and S. Li. Realistic 3D Facial Wrinkles Simulation Based on Tessellation. In *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*, volume 1, pages 250–254, Dec. 2016. doi: 10.1109/ISCID.2016.1064.

[48] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics*, 34(4):1–9, July 2015. doi: 10.1145/2766943.

[49] K. Nagano, J. Seo, J. Xing, et al. paGAN: Real-time avatars using dynamic textures. *ACM Transactions on Graphics*, 37(6):1–12, Jan. 2019. doi: 10.1145/3272127.3275075.

[50] Q. Deng, L. Ma, A. Jin, et al. Plausible 3D Face Wrinkle Generation Using Variational Autoencoders. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2021. doi: 10.1109/TVCG.2021.3051251.

[51] Blender. `https://www.blender.org/`. Accessed: 2021-09-30.

[52] Z. Zivkovic and F. van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27(7):773–780, 2006. doi: 10.1016/j.patrec.2005.11.005.

[53] E. Wood, T. Baltrusaitis, C. Hewitt, et al. 3d face reconstruction with dense landmarks, 2022.

[54] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[55] W. Wu, C. Qian, S. Yang, et al. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2138, 2018.

[56] X. Wang, L. Bo, and L. Fuxin. Adaptive wing loss for robust face alignment via heatmap

regression. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6971–6981, 2019.

[57] M. Zhu, D. Shi, M. Zheng, and M. Sadiq. Robust facial landmark detection via occlusion-adaptive deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3486–3496, 2019.

[58] B. Browatzki and C. Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *CVPR*, 2020.

[59] A. Kumar, T. K. Marks, W. Mou, et al. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8236–8246, 2020.

[60] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces In-the-wild challenge: Database and results. *Image and Vision Computing (IMAVIS)*, 2016.

[61] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2293–2303, 2019.

[62] V. F. Abrevaya, A. Boukhayma, P. H. Torr, and E. Boyer. Cross-modal deep face normals with deactivable skip connections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4979–4989, 2020.

[63] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6296–6305, 2018.

[64] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Appendices

4.A Illustrating Tension Parameters

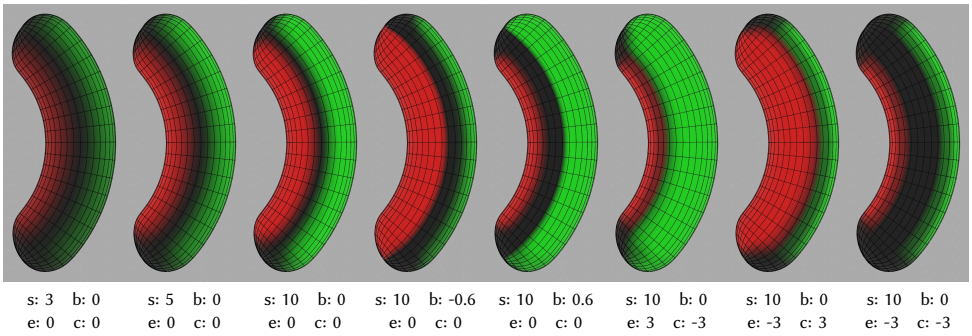


Figure 4.13: Tension Parameters - Cylinder. Illustrating the effect of varying tension parameters on a simple cylinder mesh. Legend: *s* - strength, *b* - bias, *e* - iterations for dilating/eroding expansion, *c* - iterations for dilating/eroding compression.

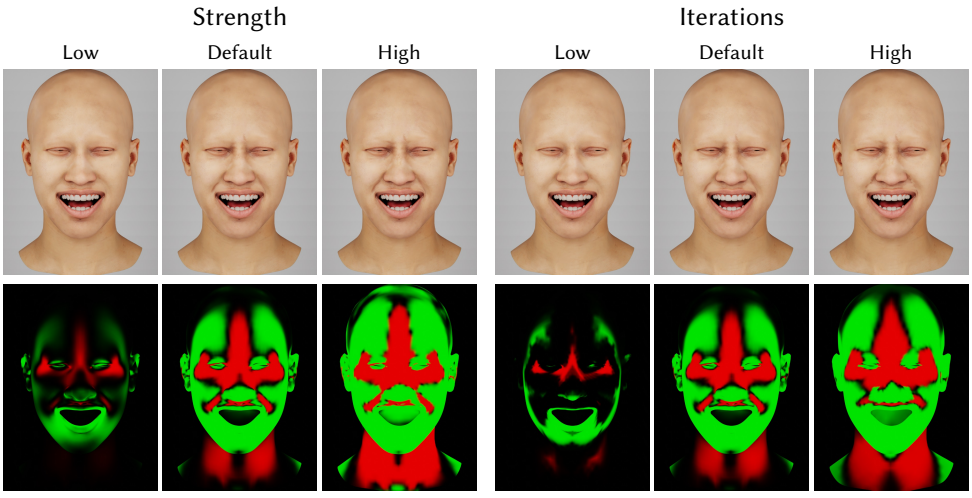


Figure 4.14: Tension Parameters - Face. Illustrating the effect of varying tension parameters on a face mesh.

4.B Eye-Region Landmark Metrics

To deal with different landmark annotation conventions (e.g. 68, 98, 703 landmarks), we use a point to polyline distance. For each eyelid point in the prediction, we measure its distance to the relevant polyline, e.g. for a predicted point on upper-left eyelid we measure

the distance from it to upper-left eyelid polyline (illustrated in Figure 4.15). This allows us to compare models with different annotation schemes and to have a better understanding of eye region error.

In cases where we do not have landmark annotations, but we know that both eyes are closed (or a single eye is closed), we can use the eye opening/aperture error instead. This is illustrated in Figure 4.16. The limitation of this approach is that it measure the relative openness of eye only and will have a low error even if the location of eyelid is wrong (but the aperture is correct). However, in combination with other metrics it provides a good signal to how well the models deal in detecting winks and blinks.

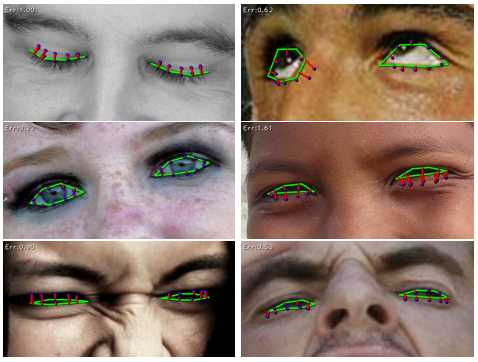


Figure 4.15: Point-to-Line Distance Metric. Green: ground truth annotation polyline. Blue: predicted eyelid landmarks. Red: residual distance. Top-left corner: associated error.

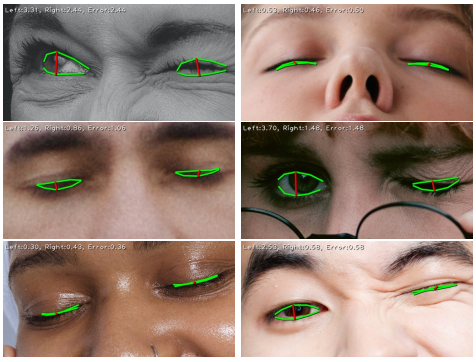


Figure 4.16: Eye Opening / Aperture Error. Green: prediction. Red: distance between the top and bottom lid that measures eye opening. Top-left corner: eye-opening of the left and right eye and the subsequent error.

4.C Landmark Predictions on 300W

We present examples of predictions on 300W dataset from models trained on real and synthetic data in Figure 4.17.

4.D Surface-Normals Predictions

In Figure 4.18 we show further comparisons to the recent face surface-normals prediction techniques of Abrevaya et al. [62] and Sengupta et al. [63].

Figure 4.19 shows failure cases from Abrevaya et al. [62] and our results on the same images. It is clear that our technique results in a significantly more robust model which can deal better with extreme lighting conditions, occlusions and darker skin tones. Note that when training our model we take the surface of glasses lenses into account, though it is also possible to ignore these and predict for the face underneath depending on how the

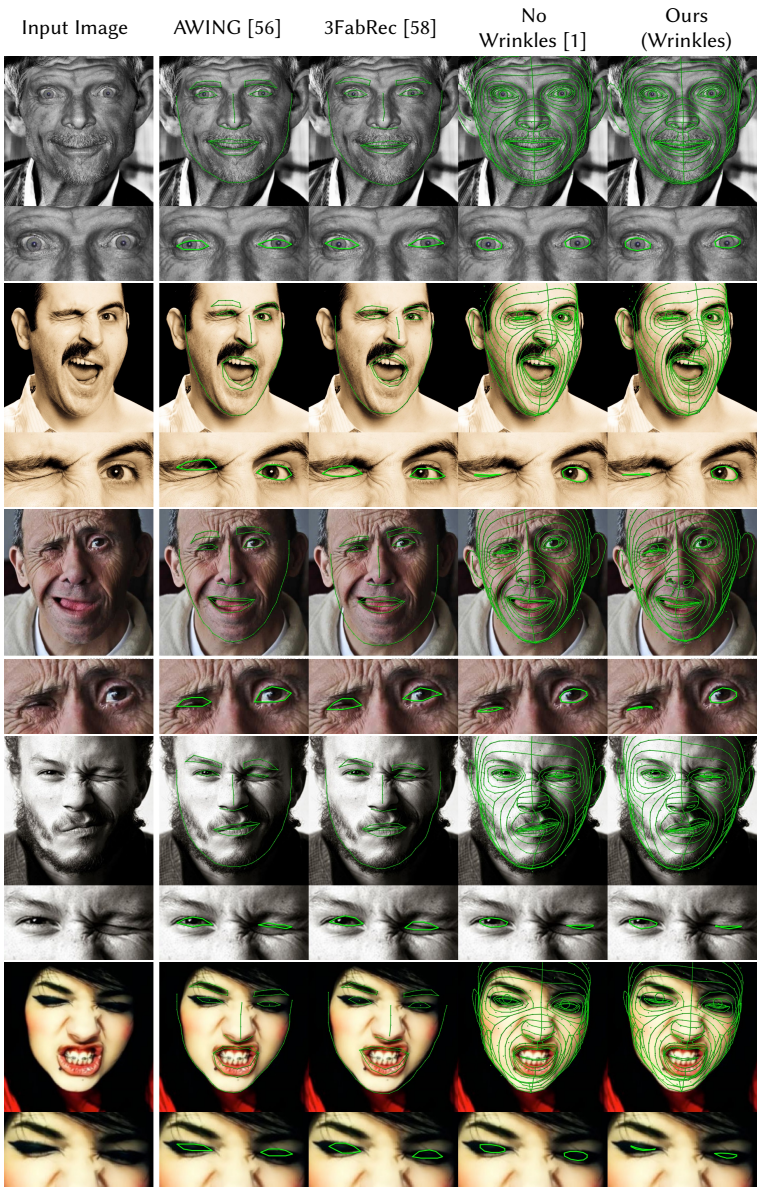


Figure 4.17: Qualitative Comparison against SOTA. Comparing prediction on 300W against SOTA models for facial landmark detection, our synthetic-only model often results in better accuracy for eye region, with improved performance for wink detection with mesh-based tension data.

rendering pipeline is configured. In all figures relating to surface-normals prediction we use our own face alignment to select the region of interest (ROI) to input to the normals

prediction U-Net model, which causes some misalignment with the ROI used in other techniques.

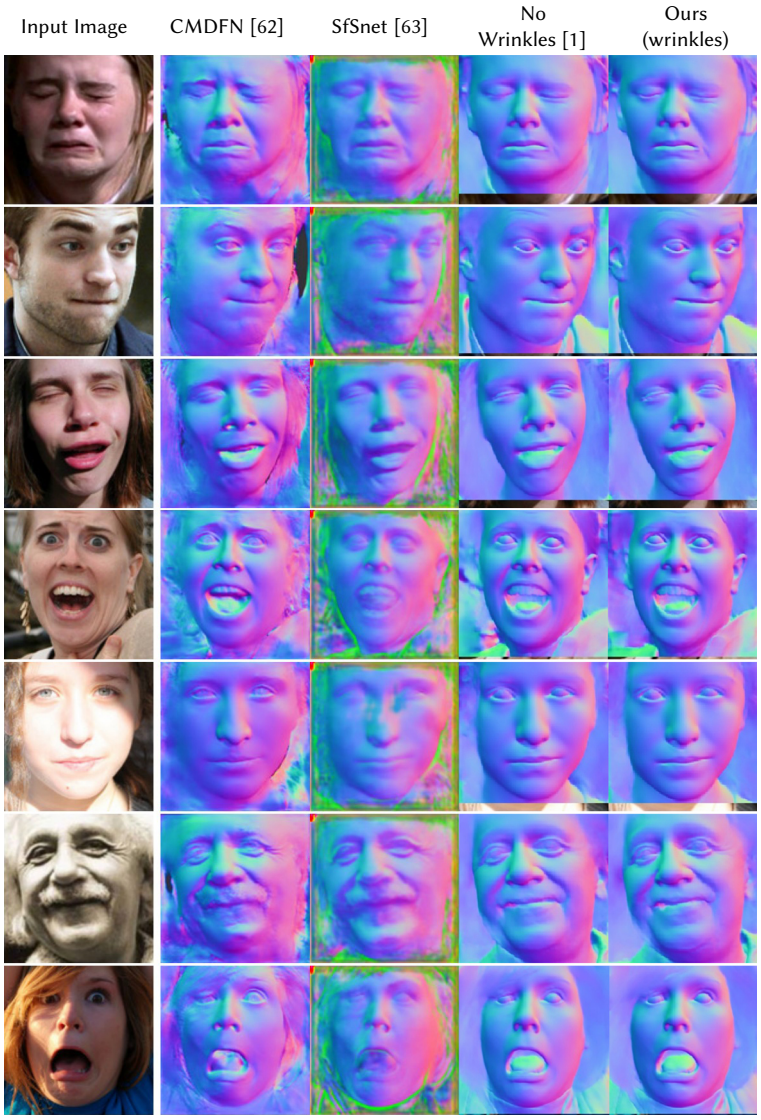


Figure 4.18: Surface-Normals Comparison. Qualitative results for surface-normals compared with two recent approaches. Note that we use our own face-alignment resulting in slight offset of the predicted ROI.

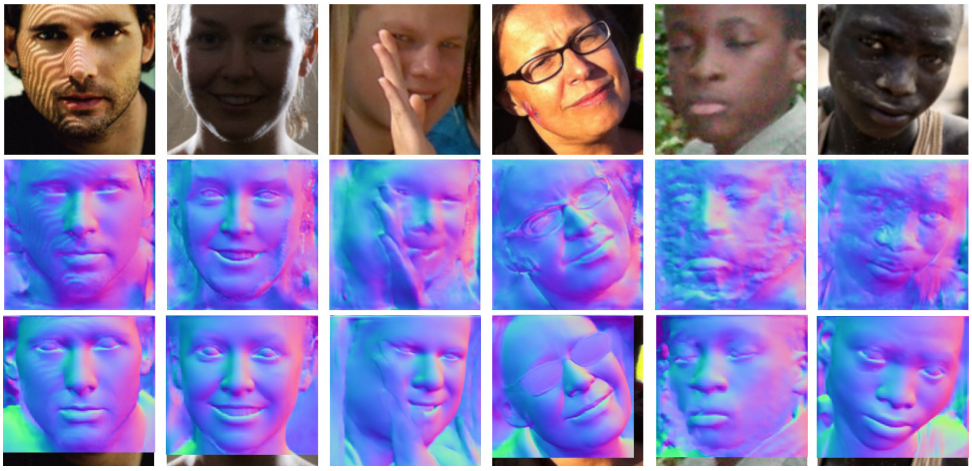


Figure 4.19: Surface-Normals Robustness. Comparison of surface-normals prediction to failure cases from Abrevaya et al. [62], showing input (top), Abrevaya et al. (middle) and Ours (bottom). Our approach is significantly more robust in cases of extreme lighting, occlusion and darker skin.

4.E The 300W-winks Subset

The subset of 300W images that make up 300W-winks is:

indoor_048, indoor_052, indoor_053, indoor_054, indoor_055, indoor_089, indoor_094, indoor_099, indoor_180, indoor_226, indoor_242, indoor_253, indoor_264, indoor_267, indoor_278, indoor_280, indoor_282, indoor_286, outdoor_073, outdoor_076, outdoor_077, outdoor_089, outdoor_097, outdoor_145, outdoor_165, outdoor_209, outdoor_243, outdoor_249, outdoor_251, outdoor_292

4.F The Pexels Winks and Blinks Dataset

All images can be accessed by appending to https://www.pexels.com/photo/

4.F.1 Blinks subset

abhishek-sinari-2026945, adrienne-andersen-2552127, adrienne-andersen-2552131, alekke-blazhin-7448048, alekke-blazhin-8450287, alekke-blazhin-8450288, alekke-blazhin-8450290, alekke-blazhin-8450296, alena-darmel-6940463, alena-shekhovtcova-7036537, alesia-kozik-7295537, alex-green-6626087, alexander-krivitskiy-4383786, alexander-stemplewski-2906663, alexandr-podvalny-1540408, alexandra-patrusheva-6806789, ali-karimiboroujeni-11381826, alina-blumberg-6925493, alyona-pastukhova-11495069, amar-preciado-10820282, amr-osman-10665375, anastasia-ilinamakarova-10832112, anastasia-shuraeva-7539962, anastasia-trofimczyk-10311002, anastasiia-chaikovska-11834502, anastasiia-shevchenko-10568846, andre-porto-7753232, andrea-gulotta-11140482, andrea-piacquadio-3757942, andrea-piacquadio-3760262, andrea-piacquadio-3760611, andrea-piacquadio-3764535, andrea-piacquadio-3768163, andrea-piacquadio-3768724, andrea-piacquadio-3771813, andrea-piacquadio-3786522, andrea-piacquadio-3799096, andrea-piacquadio-3799787, andrea-piacquadio-3799830, andrea-piacquadio-3807762, andrea-piacquadio-3811603, andrea-piacquadio-3811663, andrea-piacquadio-3812746, andrea-piacquadio-3831645, andrea-piacquadio-941693, andres-

ayrton-6578880, anete-lusina-4793357, anete-lusina-5723189, angela-roma-7479819, angelica-reyn-11893387, anh-tuan-9889769, anna-shvets-3746281, anna-shvets-3852192, anna-shvets-4557467, anna-shvets-4611655, anna-shvets-4971107, anna-shvets-5034475, anna-shvets-5069470, anna-shvets-5069493, anna-shvets-5069609, anna-tarazevich-5155727, anna-zaykina-8452431, antoni-shkraba-5890702, antoni-shkraba-7484863, arianna-jade-2896823, arina-krasnikova-6663361, arina-krasnikova-6663367, arina-krasnikova-6914826, arina-krasnikova-6914833, arina-krasnikova-6998572, arina-krasnikova-7752573, arina-krasnikova-7752693, armin-rimoldi-5269495, arsham-haghani-3423024, artyom-malyukov-11896104, azraq-al-rezoan-11763863, azraq-al-rezoan-11763868, barathan-amuthan-2723624, ben-mack-6775289, blue-bird-7210441, breno-santos-10060165, brett-sayles-4095246, caique-araujo-10218049, camilla-gari-10306657, charles-wundengba-3609781, cliff-booth-4057336, cottonbro-10049355, cottonbro-10140838, cottonbro-10678800, cottonbro-4727484, cottonbro-5020308, cottonbro-5386370, cottonbro-5561559, cottonbro-5561563, cottonbro-5850831, cottonbro-5976145, cottonbro-6700116, cottonbro-6700119, cottonbro-6700142, cottonbro-6700144, cottonbro-6753360, cottonbro-6753370, cottonbro-6753371, cottonbro-7407129, cottonbro-8102360, cottonbro-8142260, cottonbro-9063608, cottonbro-9063624, cottonbro-9063626, cottonbro-9316296, cottonbro-9467199, cottonbro-9577189, cottonbro-9955927, craig-adderley-2306203, craig-adderley-2306210, craig-adderley-2306213, cup-of-couple-6634443, cup-of-couple-6962575, cup-of-couple-6963527, daria-nekipelova-9665517, daria-rem-1977055, darina-belonogova-7886748, darina-belonogova-8386475, davner-toledo-4574403, dziana-hasanbekava-6851631, efigie-lima-marcos-11831324, ehsan-7538807, ekaterina-bolovtsova-7113346, ekaterina-bolovtsova-7113362, elina-fairytale-3865731, elina-fairytale-3865763, elina-fairytale-3865765, eman-genatilan-5348809, eman-genatilan-8589781, emmy-pua-10196907, engin-akyurt-5059305, eric-deine-11781294, estelle-umaes-11734787, evelina-zhu-6286063, faruk-tokluoglu-8777603, fireberrytech-6683091, flint-huynh-11804619, gary-barnes-6248993, gary-barnes-6249024, greta-hoffman-7675722, guilherme-almeida-1858175, hebert-santos-5485599, ichad-windhiagiri-7616249, imad-clicks-11742222, imad-clicks-11742223, ivan-mudruk-10400317, ivan-samkov-6968814, ivan-samkov-8952728, jamie-saw-10029674, jeandaniel-francoeur-7678688, jennifer-enujiugha-1904674, jill-burrow-6758033, joao-vitor-heinrichs-1787039, joshua-abner-3605015, joshua-mcknight-3290242, julia-avamotive-1070967, julia-tatyanenko-11855943, juliana-marinina-9957288, kampus-production-6298293, kampus-production-6298321, kampus-production-7928134, kampus-production-8871412, karolina-grabowska-4378486, karolina-grabowska-4498195, kat-smith-568021, ketut-subiyanto-4473864, ketut-subiyanto-4545165, ketut-subiyanto-4584390, ketut-subiyanto-4584601, kindel-media-7298396, kindel-media-7298459, kindel-media-7938549, kirill-palii-3545783, klaus-nielsen-6303717, korede-adenola-11785507, kseniya-buraya-10008858, kwesiblaq-10986569, leah-kelley-3722162, leo-acevedo-3261142, lucas-souza-1964442, lucas-souza-3608010, maksim-goncharenok-4892914, marcelo-chagas-2535859, maria-eduarda-loura-magalhaes-4340053, maria-luiza-melo-11819746, maria-orlova-4946635, maria-orlova-4947740, marija-7737766, marlon-schmeiski-11193234, mart-production-7880131, matheus-bertelli-11749497, matheus-ferrero-11470717, matheus-henrin-11360455, maycon-marmo-4346013, michelle-leman-6774345, mike-cabugao-8503888, mikhail-nilov-6707031, mikhail-nilov-6943956, mikhail-nilov-6945088, mikhail-nilov-6968191, mikhail-nilov-6968331, mikhail-nilov-7776528, mikhail-nilov-8343016, mikhail-nilov-8350479, ming-zimik-5861623, miriam-alonso-7623727, monica-turlui-8218377, monstera-5063295, monstera-5273734, monstera-5302897, monstera-5384518, monstera-6781240, monstera-6973715, monstera-6974031, monstera-6977869, monstera-7352909, mosei-films-9209576, nadin-sh-11872307, nguyen-phuong-linh-6211165, nicola-barts-7925781, nikita-nikitin-11008044, nikita-semezhin-9787604, oleg-magni-1669154, olia-danilevich-8964938, olya-prutskova-7179057, orione-conceicao-2983464, ozan-culha-11850759, ozan-culha-11858978, ozan-culha-11866492, pavel-danilyuk-7267691, pavel-danilyuk-7267700, pavel-danilyuk-7406040, pexels-user-9281097, pnw-production-8980983, pnw-production-8981313, polina-chistyakova-9052464, polina-kovaleva-5885655, polina-kovaleva-7090394, polina-tankilevitch-6630835, polina-tankilevitch-8210939, rachel-claire-4992586, rafael-freire-5714746, rafael-portraits-

9281360, raquel-silva-11870922, renthel-cueto-11131698, renthel-cueto-11131703, rfstudio-3843292, rheyan-glenn-dela-cruz-manggob-10210334, rodnae-productions-7402945, rodnae-productions-8173525, rodnae-productions-8173543, roman-odintsov-11760366, roman-odintsov-11760376, roman-odintsov-11760378, roman-odintsov-8018975, ron-lach-10139616, ron-lach-10321431, ron-lach-8159655, run-ffwpu-11757051, ruslan-rozanov-11585357, samer-daboul-4506967, samson-katt-5256085, santiago-jose-calvo-11757764, sasha-lazarev-3578326, shiny-diamond-3762659, shotpot-6338298, shvets-production-6974955, shvets-production-6975262, shvets-production-6975383, shvets-production-6975413, shvets-production-6984635, shvets-production-8005151, si-luan-pham-8778439, sound-on-3756943, svetlana-10311383, taina-bernard-3482526, tanya-gorelova-3855199, thiago-alencar-10154765, thiago-matos-10359136, thirdman-6958390, thirdman-7237074, thirdman-7268229, thirdman-7268234, thirdman-7268483, thirdman-8053704, thomas-nguka-10163670, thomas-nguka-7562643, tieu-bao-truong-8298108, tiffany-freeman-11038435, tima-miroshnichenko-5118496, tima-miroshnichenko-6670752, tubarones-photography-2737046, tubarones-photography-2943689, tubarones-photography-3065450, valdemar-9546870, vanessa-loring-5082946, vika-kirillova-10119334, vika-kirillova-11067905, vinicius-altava-2657594, vitoria-santos-1913161, vitoria-santos-2838831, vlada-karpovich-8528898, vlada-karpovich-8939842, vladimir-konoplev-11323367, vladimir-konoplev-11323376, vladimir-vasilev-7640302, wesley-carvalho-4126255, yan-krukov-6617027, yan-krukov-7155545, yana-sperry-11810044, yaroslav-shuraev-6281021, zayceva-tatiana-11210581, zayceva-tatiana-11698072

4.F.2 Winks subset

airam-datoon-9637814, alena-darmel-7322312, alena-darmel-8153597, alexander-krivitskiy-6471731, alexander-krivitskiy-6828450, amina-filkins-5560027, amina-filkins-5560029, amina-filkins-5561443, amina-filkins-5561455, andrea-piacquadio-3764391, andrea-piacquadio-3777558, andrea-piacquadio-3777563, andrea-piacquadio-3778216, andrea-piacquadio-3778673, andrea-piacquadio-3779420, andrea-piacquadio-3783107, andrea-piacquadio-3979192, arina-krasnikova-6663385, beatriz-braga-10461875, bia-sousa-2191056, brianna-amick-2069008, bruno-ticianelli-1889787, chandrashekar-hosakere-matt-707449, cleyder-duque-3690938, cottonbro-5416365, cottonbro-6144420, darija-shelkovich-5010665, dazzle-jam-2020992, debendra-das-5429301, deepak-digwal-3577006, ekaterina-glushenko-8993965, evgeny-zuchman-9986411, furkanfdemir-7020329, ganeshbabu-arun-580012, gustavo-fring-4254148, gustavo-fring-6050330, gustavo-fring-7447022, ike-louie-natividad-3208616, jekaterina-glushenko-8993965, jevgenija-shuhman-9986411, john-valette-10785650, julia-larson-6113247, julia-larson-6113626, julia-larson-6113631, julia-larson-6113639, julia-larson-6113641, kaushal-moradiya-3400573, kenneth-gorzal-surillo-8177834, ketut-subiyanto-4584180, ketut-subiyanto-4584197, ketut-subiyanto-4584387, ketut-subiyanto-4584604, koolshooters-6976854, koolshooters-7142955, koolshooters-7143075, koolshooters-7143155, koolshooters-7143217, ksenia-chernaya-6616201, laura-tancredi-7065531, matheus-bertelli-7510832, melissa-jansen-van-rensburg-2255462, mikhail-nilov-8922330, mikhail-nilov-8922333, mikhail-nilov-8923584, monstera-7114452, monstera-7114634, monstera-7139819, nataliya-vaitkevich-4813824, nataliya-vaitkevich-4813859, nichole-sebastian-3264351, nikita-saif-11494236, pavel-danilyuk-8422489, pavel-danilyuk-8422507, pavel-danilyuk-8422515, pixabay-40565, polina-tankilevitch-4723521, polina-tankilevitch-4723538, polina-tankilevitch-4725082, polina-tankilevitch-4725084, polina-tankilevitch-4725108, polina-tankilevitch-4725153, polina-tankilevitch-6630839, polina-tankilevitch-6988592, polina-zimmerman-3958873, rahmi-aksoz-9957220, rich-ortiz-5661730, rodnae-productions-10503462, rodnae-productions-6709127, ron-lach-8159657, ron-lach-8989996, sarah-chai-7262397, shvets-production-6975619, shvets-production-7525145, sora-shimazaki-5938614, th-team-7516292, the-weddingfog-9084064, thirdman-6109560, victoria-strelkaph-11034423, yan-krukov-4964933, yan-krukov-4964936, yan-krukov-4964937, yan-krukov-7793112, yuliya-shabliy-388517, zura-modebadze-4922053, zura-modebadze-4922064

II

MODELING

Forecasting and Explaining Social Behavior

5

Adaptive Forecasting of Social Cues in Conversing Groups

C. Raman, H. Hung, and M. Loog. Social Processes: Self-Supervised Meta-Learning over Conversational Groups for Forecasting Nonverbal Social Cues. *European Conference on Computer Vision (ECCV) Workshop on Computer Vision for Metaverse (cv4metaverse)*, Tel Aviv, Israel, 2022.

Abstract

Free-standing social conversations constitute a yet underexplored setting for human behavior forecasting. While the task of predicting pedestrian trajectories has received much recent attention, an intrinsic difference between these settings is how groups form and disband. Evidence from social psychology suggests that group members in a conversation explicitly self-organize to sustain the interaction by adapting to one another's behaviors. Crucially, the same individual is unlikely to adapt similarly across different groups; contextual factors such as perceived relationships, attraction, rapport, etc., influence the entire spectrum of participants' behaviors. A question arises: how can we jointly forecast the mutually dependent futures of conversation partners by modeling the dynamics unique to every group? In this paper, we propose the Social Process (SP) models, taking a novel meta-learning and stochastic perspective of group dynamics. Training group-specific forecasting models hinders generalization to unseen groups and is challenging given limited conversation data. In contrast, our SP models treat interaction sequences from a single group as a meta-dataset: we condition forecasts for a sequence from a given group on other observed-future sequence pairs from the same group. In this way, an SP model learns to adapt its forecasts to the unique dynamics of the interacting partners, generalizing to unseen groups in a data-efficient manner. Additionally, we first rethink the task formulation itself, motivating task requirements from social science literature that prior formulations have overlooked. For our formulation of Social Cue Forecasting, we evaluate the empirical performance of our SP models against both non-meta-learning and meta-learning approaches with similar assumptions. The SP models yield improved performance on synthetic and real-world behavior datasets.

Keywords: *Social Interactions, Nonverbal Cues, Behavior Forecasting*

5.1 Introduction

Picture a conversing group of people in a free-standing social setting. To conduct such exchanges, we transfer high-order social signals across space and time through explicit low-level behavior cues—examples include our pose, gestures, gaze, and floor control actions [1–3]. Evidence suggests that we employ anticipation of these and other cues to navigate daily social interactions [1, 4–8]. Consequently, for machines to truly develop adaptive social skills, they need to have the ability to forecast the future. For instance, foreseeing the upcoming behaviors of partners in advance can enable interactive agents to choose more fluid interaction policies [9], or contend with uncertainties in imperfect real-time inferences surrounding cues [3].

In literature, behavior forecasting works mainly consider data at two representations with an increasing level of abstraction: low-level cues or features that are extracted manually or automatically from raw audiovisual data, and manually labeled high-order events or

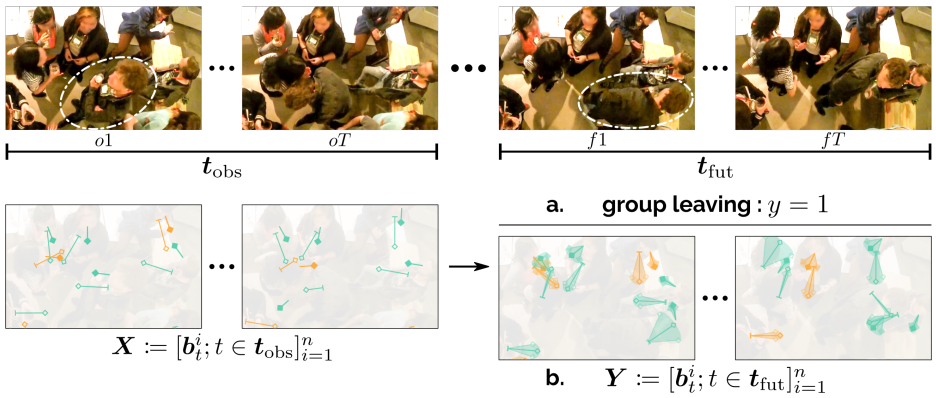


Figure 5.1: Conceptual illustration of forecasting approaches on an in-the-wild conversation from the MatchNMingle dataset [16]. **Top.** A *group leaving* event [10]: the circled individual has moved from one group in the observed window $t_{obs} = [o1 \dots oT]$ to another in a future window $t_{fut} = [f1 \dots fT]$. **Bottom.** Input behavioral cues b_t^i: head pose (solid normal), body pose (hollow normal), and speaking status (speaker in orange). **a.** The top-down approach entails predicting the event label from such cues over t_{obs}, from only 200 instances of group leaving in over 90 minutes of interaction [10]. **b.** Our proposed bottom-up, self-supervised formulation of *Social Cue Forecasting* involves regressing a future distribution for the same low-level input cues over t_{fut} (shaded spread). This enables utilizing the full 90 minutes of event-unlabeled data.

actions. The forecasting task has primarily been formulated to predict future event or action labels from observed cues or other high-order event or action labels [5, 6, 9–13]. Moreover, identifying patterns predictive of certain semantic events has been a long-standing topic of focus in the social sciences, where researchers primarily employ a top-down workflow. First, the events of interest are selected for consideration. Then their relationship to preceding cues or other high-order actions are studied in isolation through exploratory or confirmatory analysis [14, 15]. Examples of such semantic events include speaker turn transitions [5, 6], mimicry episodes [13], the termination of an interaction [9, 10], or high-order social actions [11, 12].

One hurdle in such a top-down paradigm is data efficiency. The labeled events often occur infrequently over the interaction, reducing the effective amount of labeled data. This, combined with the fact that collecting behavior data is cost and labor-intensive, precludes the effective application of neural supervised learning techniques that tend to be data demanding. More recently, some approaches have adopted a more bottom-up formulation for dyadic conversations. The task entails predicting event-independent future cues for a single target participant or virtual avatar from the preceding observed cues of both participants [17, 18]. Since training sequences are not limited to windows around semantic events, such a formulation is more data-efficient. Figure 5.1 illustrates the top-down and bottom-up approaches conceptually.

In practice, however, the concrete formulations within the bottom-up paradigm [17, 18] suffer from several conceptual problems: (i) predictions are made for a single individual using cues from both individuals as input; since people behave differently, this entails training one forecasting model per person; (ii) even so, predicting a future for one individual at a time is undesirable as these futures are not independent; and (iii) the prediction is only a single future, despite evidence that the future is not deterministic, and the same observed sequence can result in multiple socially-valid continuations [19–21].

To address all these issues, we introduce a self-supervised forecasting task called Social Cue Forecasting: predicting a *distribution* over future multimodal cues *jointly for all group members* from their same preceding multimodal cues. Note that we use *self-supervised* here to simply distinguish from the formulations where the predicted quantity (e.g. event-labels) is of a different representation than the observed input (e.g. cues). Given the cue data, the inputs and outputs of our formulation are both cues, so we *obtain the supervisory signal from the data itself.*

Furthermore, a crucial characteristic of free-standing conversations is that people sustain the interaction by explicitly adapting to one another's behaviors [1]. Moreover, the way a person adapts to their partners is a function of several complex factors surrounding their interpersonal relationships and the social setting [22, Chap. 1; 1, p. 237]. The social dynamics guiding such behavior are embedded in the constellation of participant cues and are distinct for every unique grouping of individuals. As such, a model should adapt its forecasts to the group under consideration. (Even in the pedestrian setting where coordination is only implicit, Rudenko et al. [23, Sec. 8.4.1] observe that failing to adapt predictions to different individuals is still a limitation). For our methodological contribution, we propose the probabilistic Social Processes models, viewing each conversation group as a meta-learning *task*. This allows for capturing social dynamics unique to each group without learning group-specific models and generalizing to unseen groups at evaluation in a data-efficient manner. We believe that this framing of SCF as a *few-shot* function estimation problem is especially suitable for conversation forecasting—a limited data regime where good uncertainty estimates are desirable. Concretely, we make the following contributions:

- We introduce and formalize the novel task of Social Cue Forecasting (SCF), addressing the conceptual drawbacks of past formulations.
- For SCF, we propose and evaluate the family of socially aware probabilistic Seq2Seq models we call Social Processes (SP).

5.2 Related Work

To aid readers from different disciplines situate our work within the broader research landscape, we categorize behavior-forecasting literature by interaction focus [24]. In a

focused interaction, such as conversations, participants explicitly coordinate their behaviors to sustain the interaction. In unfocused interactions, coordination is implicit, such as when pedestrians avoid collisions.

Focused Interactions. The predominant interest in conversation forecasting stems from the social sciences, with a focus on identifying patterns that are predictive of upcoming speaking turns [5–8], disengagement from an interaction [9, 10], or the splitting or merging of groups [25]. Other works forecast the time-evolving size of a group [26] or semantic social action labels [11, 12]. More recently, there has also been a growing interest in the computer vision community for tasks related to inferring low-level cues of participants either from their partners' cues [27] or raw multimodal sensor data [28]. Here there has also been some interest in forecasting nonverbal behavior, mainly for dyadic interactions [17, 18, 29]. The task involves forecasting the future cues of a target individual from the preceding cues of both participants.

Unfocused Interactions. Early approaches for forecasting pedestrian or vehicle trajectories were heuristic-based, involving hand-crafted energy potentials to describe the influence pedestrians and vehicles have on each other [30–37]. Recent approaches build upon the idea of encoding relative positional information directly into a neural architecture [38–45]. Some works go beyond locations, predicting keypoints in group activities [46, 47]. Rudenko et al. [23] provide a survey of approaches within this space.

Non-Interaction Settings. Here, the focus has been on forecasting individual poses from images [48] and video [49, 50], or synthesizing poses using high-level control parameters [51, 52]. The self-supervised aspects of our task formulation are related to visual forecasting, where the goal has been to predict non-semantic low-level pixel features or intermediate representations [34, 50, 53–57]. Such learned representations have been utilized for other tasks like semi-supervised classification [58], or training agents in immersive environments [59].

For the interested reader, we further discuss practical considerations distinguishing forecasting in conversation and pedestrian settings in Appendix 5.E.

5.3 Social Cue Forecasting: Task Formalization

While self-supervision has shown promise for learning representations of language and video data, is this bottom-up approach conceptually reasonable for behavior cues? The crucial observation we make is that the semantic meaning transferred in interactions (the so-called *social signal* [60]) is already embedded in the low-level cues [61]. So representations of this high-level semantic meaning that we associate with actions and events (e.g. *group leaving*) can be learned from the low-level dynamics in the cues.

5.3.1 Formalization and Distinction from Prior Task Formulations

The objective of SCF is to predict future behavioral cues of *all* people involved in a social encounter given an observed sequence of their behavioral features. Formally, let us denote a window of monotonically increasing observed timesteps as $\boldsymbol{t}_{\text{obs}} = [o1, o2, ..., oT]$, and an unobserved future time window as $\boldsymbol{t}_{\text{fut}} = [f1, f2, ..., fT]$, $f1 > oT$. Note that $\boldsymbol{t}_{\text{fut}}$ and $\boldsymbol{t}_{\text{obs}}$ can be of different lengths, and $\boldsymbol{t}_{\text{fut}}$ need not immediately follow $\boldsymbol{t}_{\text{obs}}$. Given n interacting participants, let us denote their social cues over $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ as

$$\boldsymbol{X} = [\boldsymbol{b}_t^i; t \in \boldsymbol{t}_{\text{obs}}]_{i=1}^n, \quad \boldsymbol{Y} = [\boldsymbol{b}_t^i; t \in \boldsymbol{t}_{\text{fut}}]_{i=1}^n. \tag{5.1a, b}$$

The vector $\boldsymbol{b}_t^i$ encapsulates the multimodal cues of interest from participant i at time t. These can include head and body pose, speaking status, facial expressions, gestures, verbal content—any information streams that combine to transfer social meaning.

5

Distribution over Futures. In its simplest form, given an $\boldsymbol{X}$, the objective of SCF is to learn a single function f such that $\boldsymbol{Y} = f(\boldsymbol{X})$. However, an inherent challenge in forecasting behavior is that an observed sequence of interaction does not have a deterministic future and can result in multiple socially valid ones—a window of overlapping speech between people may and may not result in a change of speaker [19, 20], a change in head orientation may continue into a sweeping glance across the room or a darting glance stopping at a recipient of interest [21]. In some cases, certain observed behaviors—intonation and gaze cues [5, 62] or synchronization in speaker-listener speech [63] for turn-taking—may make some outcomes more likely than others. Given that there are both supporting and challenging arguments for how these observations influence subsequent behaviors [63, p. 5; 62, p. 22], it would be beneficial if a data-driven model expresses a measure of uncertainty in its forecasts. We do this by modeling the distribution over possible futures $p(\boldsymbol{Y}|\boldsymbol{X})$, rather than a single future $\boldsymbol{Y}$ for a given $\boldsymbol{X}$, the latter being the case for previous formulations for cues [18, 27, 46] and actions [11, 12].

Joint Modeling of Future Uncertainty. A defining characteristic of focused interactions is that the participants sustain the shared interaction through explicit, cooperative coordination of behavior [1, p. 220]—the futures of interacting individuals are not independent given an observed window of group behavior. It is therefore essential to capture uncertainty in forecasts at the *global* level—jointly forecasting one future for all participants at a time, rather than at a *local* output level—one future for each individual independent of the remaining participants' futures. In contrast, applying the prior formulations [17, 18, 27] requires the training of separate models treating each individual as a target (for the same group input) and then forecasting an independent future one at a time. Meanwhile, other

prior pose forecasting works [48–52] have been in non-social settings and do not need to model such behavioral interdependence.

Non-Contiguous Observed and Future Windows. Domain experts are often interested in settings where $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ are offset by an arbitrary delay, such as forecasting a time lagged synchrony [64] or mimicry [13] episode, or upcoming disengagement [9, 10]. We therefore allow for non-contiguous $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$. Operationalizing prior formulations that predict one step into the future [11, 12, 27, 46] would entail a sliding window of autoregressive predictions over the offset between $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ (from oT to $f1$), with errors cascading even before decoding is performed over the window of interest $\boldsymbol{t}_{\text{fut}}$.

Our task formalization of SCF can be viewed as a social science-grounded generalization of prior computational formulations, and therefore suitable for a wider range of cross-disciplinary tasks, both computational and analytical.

5.4 Method Preliminaries

Meta-Learning. A supervised learning algorithm can be viewed as a function mapping a dataset $C \coloneqq (\boldsymbol{X}_C, \boldsymbol{Y}_C) \coloneqq \{(\boldsymbol{x}^i, \boldsymbol{y}^i)\}_{i \in [N_C]}$ to a predictor $f(\boldsymbol{x})$. Here N_C is the number of datapoints in C, and $[N_C] \coloneqq \{1, \ldots, N_C\}$. The key idea of meta-learning is to learn how to learn from a dataset in order to adapt to unseen supervised tasks; hence the name *meta*-learning. This is done by learning a map $C \mapsto f(\cdot, C)$. In meta-learning literature, a *task* refers to each dataset in a collection $\{\mathcal{T}_m\}_{m=1}^{N_{\text{tasks}}}$ of related datasets [65]. Training is episodic, where each task $\mathcal{T}$ is split into subsets (C, D). A meta-learner then fits the subset of target points D given the subset of context observations C. At meta-test time, the resulting predictor $f(\boldsymbol{x}, C)$ is adapted to make predictions for target points on an unseen task by conditioning on a new context set C unseen during meta-training.

Neural Processes (NPs). Sharing the same core motivations, NPs [66] can be viewed as a family of latent variable models that extend the idea of meta-learning to situations where uncertainty in the predictions $f(\boldsymbol{x}, C)$ are desirable. They do this by meta-learning a map from datasets to stochastic processes, estimating a distribution over the predictions $p(\boldsymbol{Y}|\boldsymbol{X}, C)$. To capture this distribution, NPs model the conditional latent distribution $p(\boldsymbol{z}|C)$ from which a task representation $\boldsymbol{z} \in \mathbb{R}^d$ is sampled. This introduces stochasticity, constituting what is called the model's *latent path*. The context can also be directly incorporated through a *deterministic path*, via a representation $\boldsymbol{r}_C \in \mathbb{R}^d$ aggregated over C. An observation model $p(\boldsymbol{y}^i|\boldsymbol{x}^i, \boldsymbol{r}_C, \boldsymbol{z})$ then fits the target observations in D. The generative process for the NP is written as

$$p(\boldsymbol{Y}|\boldsymbol{X}, C) \coloneqq \int p(\boldsymbol{Y}|\boldsymbol{X}, C, \boldsymbol{z}) p(\boldsymbol{z}|C) d\boldsymbol{z} = \int p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{r}_C, \boldsymbol{z}) q(\boldsymbol{z}|\boldsymbol{s}_C) d\boldsymbol{z}, \tag{5.2}$$

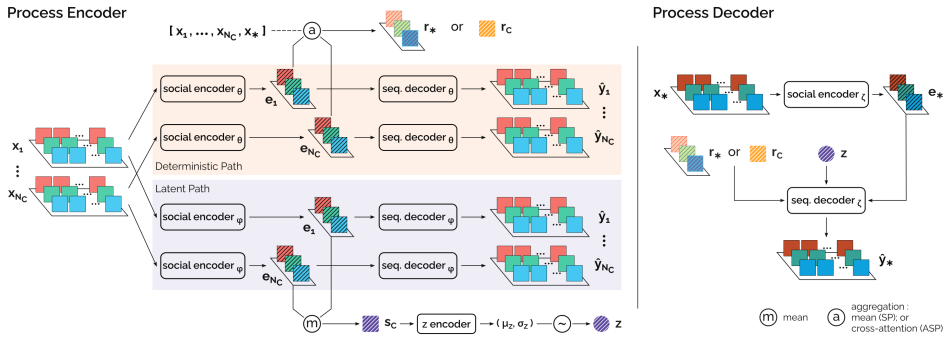


Figure 5.2: Architecture of the SP and ASP family.

where $p(Y|\boldsymbol{X}, \boldsymbol{r}_C, z) = \prod_{i\in[N_D]} p(\boldsymbol{y}^i|\boldsymbol{x}^i, \boldsymbol{r}_C, z)$. The latent z is modeled by a factorized Gaussian parameterized by $\boldsymbol{s}_C = f_s(C)$, with f_s being a deterministic function invariant to order permutation over C. When the conditioning on context is removed ($C = \emptyset$), we have $q(z|\boldsymbol{s}_\emptyset) = p(z)$, the zero-information prior on z. The deterministic path uses a function f_r similar to f_s, so that $\boldsymbol{r}_C = f_r(C)$. In practice this is implemented as $\boldsymbol{r}_C = \sum_{i\in[N_C]} \text{MLP}(\boldsymbol{x}_i, \boldsymbol{y}_i)/N_C$. The observation model is referred to as the *decoder*, and q, f_r, f_s comprise the *encoders*. The parameters of the NP are learned for random subsets C and D for a task by maximizing the evidence lower bound (ELBO)

$$\log p(Y|\boldsymbol{X}, C) \geq \mathbb{E}_{q(z|\boldsymbol{s}_D)}[\log p(Y|\boldsymbol{X}, C, z)] - \mathbb{KL}(q(z|\boldsymbol{s}_D)||q(z|\boldsymbol{s}_C)). \tag{5.3}$$

5.5 Social Processes: Methodology

Our core idea for adapting predictions to a group's unique behavioral dynamics is to condition forecasts on a context set C of the same group's observed-future sequence pairs. By *learning to learn*, i.e., *meta-learn* from a context set, our model can generalize to unseen groups at evaluation by conditioning on an unseen context set of the test group's behavior sequences. In practice, a social robot might, for instance, observe such an evaluation context set before approaching a new group.

We set up by splitting the interaction into pairs of observed and future sequences, writing the context as $C = (\boldsymbol{X}_C, \boldsymbol{Y}_C) = (\boldsymbol{X}_j, \boldsymbol{Y}_k)_{(j,k)\in[N_C]\times[N_C]}$, where every $\boldsymbol{X}_j$ occurs before the corresponding $\boldsymbol{Y}_k$. Since we allow for non-contiguous $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$, the jth $\boldsymbol{t}_{\text{obs}}$ can have multiple associated $\boldsymbol{t}_{\text{fut}}$ windows for prediction, up to a maximum offset. Denoting the set of target window pairs as $D = (\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{X}_j, \boldsymbol{Y}_k)_{(j,k)\in[N_D]\times[N_D]}$, our goal is to model the distribution $p(\boldsymbol{Y}|\boldsymbol{X}, C)$. Note that when conditioning on context is removed ($C = \emptyset$), we simply revert to the non-meta-learning formulation $p(\boldsymbol{Y}|\boldsymbol{X})$.

The generative process for our Social Process (SP) model follows Equation 5.2, which we

extend to social forecasting in two ways. We embed an observed sequence $\boldsymbol{x}^i$ for participant p_i into a condensed encoding $\boldsymbol{e}^i \in \mathbb{R}^d$ that is then decoded into the future sequence using a Seq2Seq architecture [67, 68]. Crucially, the sequence decoder only accesses $\boldsymbol{x}^i$ through $\boldsymbol{e}^i$. So after training, $\boldsymbol{e}^i$ must encode the *temporal* information that $\boldsymbol{x}^i$ contains about the future. Further, social behavior is interdependent. We model $\boldsymbol{e}^i$ as a function of both, p_i's own behavior as well as that of partners $\mathrm{p}_{j,j\neq i}$ from p_i's perspective. This captures the *spatial* influence partners have on the participant over $\boldsymbol{t}_{\text{obs}}$. Using notation we established in Section 5.3, we define the observation model for p_i as

$$p(\boldsymbol{y}^i|\boldsymbol{x}^i, C, \boldsymbol{z}) = p(\boldsymbol{b}^i_{f1}, \ldots, \boldsymbol{b}^i_{fT}|\boldsymbol{b}^i_{o1}, \ldots, \boldsymbol{b}^i_{oT}, C, \boldsymbol{z}) = p(\boldsymbol{b}^i_{f1}, \ldots, \boldsymbol{b}^i_{fT}|\boldsymbol{e}^i, \boldsymbol{r}_C, \boldsymbol{z}). \tag{5.4}$$

If decoding is carried out in an auto-regressive manner, the right hand side of Equation 5.4 simplifies to $\prod_{t=f1}^{fT} p(\boldsymbol{b}^i_t|\boldsymbol{b}^i_{t-1}, \ldots, \boldsymbol{b}^i_{f1}, \boldsymbol{e}^i, \boldsymbol{r}_C, \boldsymbol{z})$. Following the standard NP setting, we implement the observation model as a set of Gaussian distributions factorized over time and feature dimensions. We also incorporate the cross-attention mechanism from the Attentive Neural Process (ANP) [69] to define the variant Attentive Social Process (ASP). Following Equation 5.4 and the definition of the ANP, the corresponding observation model of the ASP for a single participant is defined as

$$p(\boldsymbol{y}^i|\boldsymbol{x}^i, C, \boldsymbol{z}) = p(\boldsymbol{b}^i_{f1}, \ldots, \boldsymbol{b}^i_{fT}|\boldsymbol{e}^i, r^*(C, \boldsymbol{x}^i), \boldsymbol{z}). \tag{5.5}$$

Here each target query sequence $\boldsymbol{x}^i_*$ attends to the context sequences $\boldsymbol{X}_C$ to produce a query-specific representation $\boldsymbol{r}_* = r^*(C, \boldsymbol{x}^i_*) \in \mathbb{R}^d$.

The model architectures are illustrated in Figure 5.2. Note that our modeling assumption is that the underlying stochastic process generating social behaviors does not evolve over time. That is, the individual factors determining how participants coordinate behaviors—age, cultural background, personality variables [22, Chap. 1; 1, p. 237]—are likely to remain the same over a single interaction. This is in contrast to the line of work that deals with *meta-transfer learning*, where the stochastic process itself changes over time [70–73]; this entails modeling a different $\boldsymbol{z}$ distribution for every timestep.

Encoding Partner Behavior. To encode partners' influence on an individual's future, we use a pair of sequence encoders: one to encode the temporal dynamics of participant p^i's features, $\boldsymbol{e}^i_{\text{self}} = f_{\text{self}}(\boldsymbol{x}^i)$, and another to encode the dynamics of a transformed representation of the features of p^i's partners, $\boldsymbol{e}^i_{\text{partner}} = f_{\text{partner}}(\psi(\boldsymbol{x}^{j,(j\neq i)}))$. Using a separate network to encode partner behavior enables sampling an individual's and partners' features at different sampling rates.

How do we model $\psi(\boldsymbol{x}^{j,(j\neq i)})$? We want the partners' representation to possess two properties: *permutation invariance*—changing the order of the partners should not affect the representation, and *group-size independence*—we want to compactly represent all partners

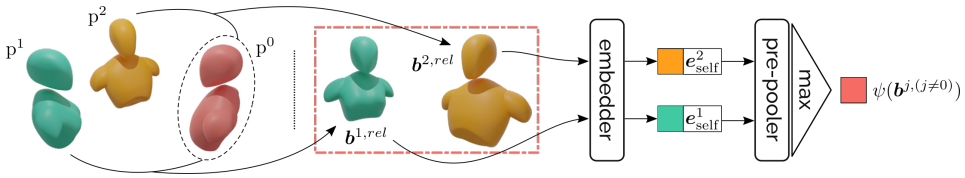


Figure 5.3: Encoding partner behavior for participant p^0 for a single timestep. To model the influence partners p^1 and p^2 have on the behavior of p^0, we transform the partner features to capture the interaction from p^0's perspective, and learn a representation of these features invariant to group size and partner-order permutation using the symmetric max function.

independent of the group size. Intuitively, to model partner influence on p^i, we wish to *capture a view of the partners' behavior as* p^i *perceives it.* Figure 5.3 illustrates the underlying intuition. We do this by computing pooled embeddings of relative behavioral features, extending Gupta et al. [40]'s approach for pedestrian positions to conversation behavior. Note that our partner-encoding approach is in contrast to that of Tan et al. [28], which is order and group-size dependent, and Yao et al. [46], who do not transform the partner features to an individual's perspective.

Since the most commonly considered cues in literature are pose (orientation and location) and binary speaking status [28, 74, 75], we specify how we transform them. For a single timestep, we denote these cues for p^i as $\boldsymbol{b}^i = [\boldsymbol{q}^i; \boldsymbol{l}^i; s^i]$, and for p^j as $\boldsymbol{b}^j = [\boldsymbol{q}^j; \boldsymbol{l}^j; s^j]$. We compute the relative partner features $\boldsymbol{b}^{j,rel} = [\boldsymbol{q}^{rel}; \boldsymbol{l}^{rel}; s^{rel}]$ by transforming $\boldsymbol{b}^j$ to a frame of reference defined by $\boldsymbol{b}^i$:

$$\boldsymbol{q}^{rel} = \boldsymbol{q}^i * (\boldsymbol{q}^j)^{-1}, \quad \boldsymbol{l}^{rel} = \boldsymbol{l}^j - \boldsymbol{l}^i, \quad s^{rel} = s^j - s^i. \tag{5.6a-c}$$

Note that we use unit quaternions (denoted $\boldsymbol{q}$) for representing orientation due to their various benefits over other representations of rotation [76, Sec. 3.2]. The operator $*$ denotes the Hamilton product of the quaternions. These transformed features $\boldsymbol{b}^{j,rel}$ for each p^j are then encoded using an *embedder* MLP. The outputs are concatenated with their corresponding $\boldsymbol{e}^j_{\text{self}}$ and processed by a *pre-pooler* MLP. Assuming d_{in} and d_{out} pre-pooler input and output dims and J partners, we stack the J inputs to obtain (J, d_{in}) tensors. The (J, d_{out})-dim output is element-wise max-pooled over the J dim, resulting in the d_{out}-dim vector $\psi(\boldsymbol{b}^{j,(j\neq i)})$ for any value of J, per timestep. We capture the temporal dynamics in this pooled representation over $\boldsymbol{t}_{\text{obs}}$ using f_{partner}. Finally, we combine $\boldsymbol{e}^i_{\text{self}}$ and $\boldsymbol{e}^i_{\text{partner}}$ for p^i through a linear projection (defined by a weight matrix W) to obtain the individual's embedding $\boldsymbol{e}^i_{\text{ind}} = W \cdot [\boldsymbol{e}^i_{\text{self}}; \boldsymbol{e}^i_{\text{partner}}]$. Our intuition is that with information about both p^i themselves, and of p^i's partners from p^i's point-of-view, $\boldsymbol{e}^i_{\text{ind}}$ now contains the information required to predict p^i's future behavior.

Encoding Future Window Offset. Since we allow for non-contiguous windows, a single $\boldsymbol{t}_{\text{obs}}$ might be associated to multiple $\boldsymbol{t}_{\text{fut}}$ windows at different offsets. Decoding the same $\boldsymbol{e}^i_{\text{ind}}$ into multiple sequences (for different $\boldsymbol{t}_{\text{fut}}$) in the absence of any timing information might cause an averaging effect in either the decoder or the information encoded in $\boldsymbol{e}^i_{\text{ind}}$. One option would be to immediately start decoding after $\boldsymbol{t}_{\text{obs}}$ and discard the predictions in the offset between $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$. However, auto-regressive decoding might lead to cascading errors over the offset. Instead, we address this one-to-many issue by injecting the offset information into $\boldsymbol{e}^i_{\text{ind}}$. The decoder then receives a unique encoded representation for every $\boldsymbol{t}_{\text{fut}}$ corresponding to the same $\boldsymbol{t}_{\text{obs}}$. We do this by repurposing the idea of sinusoidal positional encodings [77] to encode window offsets rather than relative token positions in sequences. For a given $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$, and d_e-dim $\boldsymbol{e}^i_{\text{ind}}$ we define the offset as $\Delta t = f1 - oT$, and the corresponding offset encoding $OE_{\Delta t}$ as

$$OE_{(\Delta t,2m)} = \sin(\Delta t/10000^{2m/d_e}), OE_{(\Delta t,2m+1)} = \cos(\Delta t/10000^{2m/d_e}). \tag{5.7a, b}$$

Here m refers to the dimension index in the encoding. We finally compute the representation $\boldsymbol{e}^i$ for Equation 5.4 and Equation 5.5 as

$$\boldsymbol{e}^i = \boldsymbol{e}^i_{\text{ind}} + OE_{\Delta t}. \tag{5.8}$$

Auxiliary Loss Functions. We incorporate a geometric loss function for each of our sequence decoders to improve performance in pose regression tasks. For p_i at time t, given the ground truth $\boldsymbol{b}^i_t = [\boldsymbol{q}; \boldsymbol{l}; s]$, and the predicted mean $\hat{\boldsymbol{b}}^i_t = [\hat{\boldsymbol{q}}; \hat{\boldsymbol{l}}; \hat{\mathrm{s}}]$, we denote the tuple $(\boldsymbol{b}^i_t, \hat{\boldsymbol{b}}^i_t)$ as B^i_t. We then have the location loss in Euclidean space $\mathscr{L}_{\mathrm{l}}(B^i_t) = \|\boldsymbol{l} - \hat{\boldsymbol{l}}\|$, and we can regress the quaternion values using

$$\mathscr{L}_{\mathrm{q}}(B^i_t) = \left\| \boldsymbol{q} - \frac{\hat{\boldsymbol{q}}}{\|\hat{\boldsymbol{q}}\|} \right\|. \tag{5.9}$$

Kendall and Cipolla [76] show how these losses can be combined using the homoscedastic uncertainties in position and orientation, $\hat{\sigma}^2_{\mathrm{l}}$ and $\hat{\sigma}^2_{\mathrm{q}}$:

$$\mathscr{L}_{\sigma}(B^i_t) = \mathscr{L}_{\mathrm{l}}(B^i_t) \exp(-\hat{\mathrm{s}}_{\mathrm{l}}) + \hat{\mathrm{s}}_{\mathrm{l}} + \mathscr{L}_{\mathrm{q}}(B^i_t) \exp(-\hat{\mathrm{s}}_{\mathrm{q}}) + \hat{\mathrm{s}}_{\mathrm{q}}, \tag{5.10}$$

where $\hat{s} \coloneqq \log \hat{\sigma}^2$. Using the binary cross-entropy loss for speaking status $\mathscr{L}_{\mathrm{s}}(B^i_t)$, we have the overall auxiliary loss over $t \in \boldsymbol{t}_{\text{fut}}$:

$$\mathscr{L}_{\text{aux}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_i \sum_t \mathscr{L}_{\sigma}(B^i_t) + \mathscr{L}_{\mathrm{s}}(B^i_t). \tag{5.11}$$

The parameters of the SP and ASP are trained by maximizing the ELBO (Equation 5.3) and minimizing this auxiliary loss.

5.6 Experiments and Results

5.6.1 Experimental Setup

Evaluation Metrics. Prior forecasting formulations output a single future. However, since the future is not deterministic, we predict a future distribution. Consequently, needing a metric that accounts for probabilistic predictions, we report the log-likelihood (LL) $\log p(Y|X, C)$, commonly used by all variants within the NP family [66, 69, 70]. The metric is equal to the log of the predicted density evaluated at the ground-truth value. (Note: the fact that the vast majority of forecasting works even in pedestrian settings omit a probabilistic metric, using only geometric metrics, is a limitation also observed by Rudenko et al. [23, Sec. 8.3].) Nevertheless, for additional insight beyond the LL, we also report the errors in the predicted means—geometric errors for pose and accuracy for speaking status—and provide qualitative visualizations of forecasts.

Models and Baselines. In keeping with the task requirements and for fair evaluation, we require that all models we compare against forecast a distribution over future cues.

- To evaluate our core idea of viewing conversing groups as meta-learning tasks, we compare against non-meta-learning methods: we adapt variational encoder-decoder (VED) architectures [78, 79] to output a distribution.
- To evaluate our specific modeling choices within the meta-learning family, we compare against the NP and ANP models (see Section 5.5). The original methods were not proposed for sequences, so we adapt them by collapsing the timestep and feature dimensions in the data.

Note that in contrast to the SP models, these baselines have direct access to the future sequences in the context, and therefore constitute a strong baseline. We consider two variants for both NP and SP models: *-latent* denoting only the stochastic path; and *-uniform* containing both the deterministic and stochastic paths with uniform attention over context sequences. We further consider two attention mechanisms for the cross-attention module: *-dot* with dot attention, and *-mh* with wide multi-head attention [69]. Finally, we experiment with two choices of backbone architectures: multi-layer perceptrons (MLP), and Gated Recurrent Units (GRU). Implementation and training details can be found in Appendix 5.D. Code, processed data, trained models, and test batches for reproduction are available at https://github.com/chiragraman/social-processes.

5.6.2 Evaluation on Synthesized Behavior Data

To first validate our method on a toy task, we synthesize a dataset simulating two glancing behaviors in social settings [21], approximated by horizontal head rotation. The sweeping *Type I* glance is represented by a 1D sinusoid over 20 timesteps. The gaze-fixating *Type III* glance is denoted by clipping the amplitude for the last six timesteps. The task is to forecast

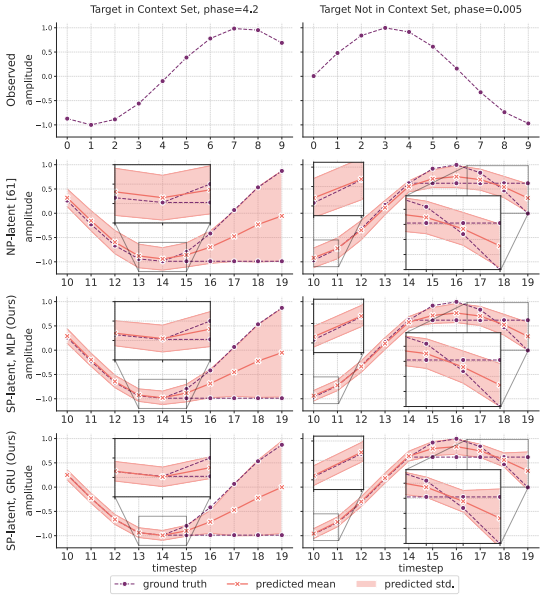


Figure 5.4: Ground truths and predictions for the toy task of forecasting simulated glancing behavior. Our SP models learn a better fit than the NP model, SP-GRU being the best (see zoomed insets).

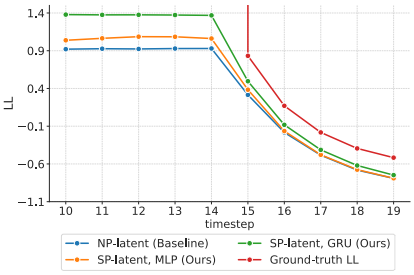


Figure 5.5: Mean per timestep LL over the sequences in the synthetic glancing dataset. Higher is better.

Table 5.1: Mean (Std.) Metrics on the Synthetic Glancing Behavior Dataset. The metrics are averaged over timesteps; mean and std. are then computed over sequences. Higher is better for LL, lower for MAE.

	LL	Head Ori. MAE (°)
NP-latent	0.28 (0.24)	19.63 (7.26)
SP-latent (MLP)	0.36 (0.20)	19.46 (7.05)
SP-latent (GRU)	**0.55 (0.23)**	**18.55 (7.11)**

the signal over the last 10 timesteps ($\boldsymbol{t}_{\text{fut}}$) by observing the first 10 ($\boldsymbol{t}_{\text{obs}}$). Consequently, the first half of $\boldsymbol{t}_{\text{fut}}$ is certain, while the last half is uncertain: every observed sinusoid has two ground truth futures in the data (clipped and unclipped). It is impossible to infer from an observed sequence alone if the head rotation will stop partway through the future. Figure 5.4 illustrates the predictions for two sample sequences. Table 5.1 provides quantitative metrics and Figure 5.5 plots the LL per timestep. The LL is expected to decrease over timesteps where ground-truth futures diverge, being ∞ when the future is certain. We observe that all models estimate the mean reasonably well, although our proposed SP models perform best. More crucially, the SP models, especially the SP-GRU, learn much better uncertainty estimates compared to the NP baseline (see zoomed regions in Figure 5.4). We provide additional analysis, alternative qualitative visualizations, and data synthesis details in Appendices 5.A, 5.B, and 5.C respectively.

5.6.3 Evaluation on Real-World Behavior Data

Datasets and Preprocessing. With limited behavioral data availability, a common practice in the domain is to solely train and evaluate methods on synthesized behavior dynamics [12, 80]. In contrast, we also evaluate on two real-world behavior datasets: the MatchN-

Table 5.2: Mean (Std.) Log-Likelihood (LL) on the MatchNMingle and Haggling Test Sets. For a single sequence, we sum over the feature and participant dimensions, and average over timesteps. The reported mean and std. are over individual sequences in the test sets. Higher is better. Underline indicates best LL within family.

	MatchNMingle		Haggling	
	Random	**Fixed-Initial**	**Random**	**Fixed-Initial**
VED Family [78, 79]				
VED-MLP	8.1 (7.2)	7.9 (7.0)	4.0 (8.3)	4.1 (8.2)
VED-GRU	25.4 (18.0)	25.1 (19.1)	60.3 (2.2)	60.3 (2.1)
NP Family [66, 69]				
NP-latent	22.1 (17.8)	21.6 (18.5)	27.2 (17.3)	27.9 (16.3)
NP-uniform	21.4 (18.8)	20.5 (17.8)	24.8 (22.9)	25.0 (22.2)
ANP-dot	22.8 (18.6)	21.0 (18.3)	26.7 (21.4)	24.7 (20.8)
ANP-mh	23.6 (15.6)	20.0 (23.9)	25.1 (23.1)	24.8 (22.4)
Ours (SP-MLP)				
SP-latent	102.1 (29.9)	101.5 (29.2)	136.6 (7.0)	136.7 (7.0)
SP-uniform	112.8 (34.1)	111.4 (33.8)	138.3 (8.0)	137.6 (8.4)
ASP-dot	109.9 (32.9)	107.6 (32.1)	137.8 (7.5)	136.4 (7.6)
ASP-mh	112.9 (34.7)	111.3 (33.6)	146.0 (10.9)	145.7 (10.2)
Ours (SP-GRU)				
SP-latent	86.4 (37.2)	85.4 (37.2)	66.7 (27.4)	66.2 (30.7)
SP-uniform	87.0 (38.4)	85.5 (38.3)	79.9 (50.5)	78.6 (52.2)
ASP-dot	87.6 (39.1)	83.9 (38.1)	38.4 (60.4)	27.2 (93.4)
ASP-mh	85.8 (37.1)	82.3 (36.0)	66.3 (30.3)	59.3 (32.4)

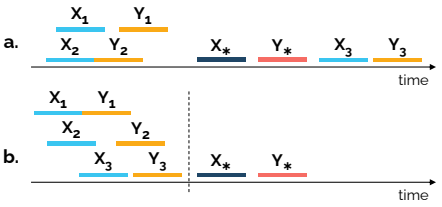


Figure 5.6: Context Regimes. For a target sequence pair (X_*, Y_*), context pairs (here 3) are sampled either **a.** randomly across the lifetime of the group interaction (*random*), or **b.** from a fixed initial duration (*fixed-initial*).

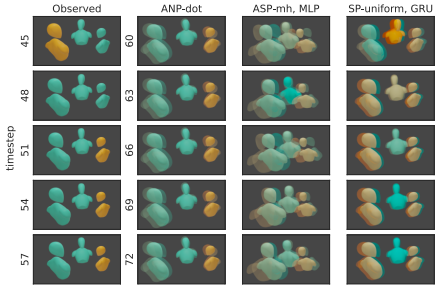


Figure 5.7: Forecasts over selected timesteps from the Haggling group *170224-a1-group1*. Speaking status is interpolated between orange (speaking) and blue (listening). Translucent models denote the predicted mean ± std.

Mingle (MnM) dataset of in-the-wild mingling behavior [16], and the Haggling dataset of a triadic game where two sellers compete to sell a fictional product to a buyer [27]. For MnM, we treat the 42 groups from Day 1 as test sets and a total of 101 groups from the other two days as train sets. For Haggling, we use the same split of 79 training and 28 test groups used by Joo et al. [27]. We consider the following cues: *head pose* and *body pose*, described by the location of a keypoint and an orientation quaternion; and binary *speaking status*. These are the most commonly considered cues in computational analyses of conversations [28, 74, 75] given how crucial they are in sustaining interactions [1, 20, 61]. For orientation, we first convert the normal vectors (provided in the horizontal direction in both datasets) into unit quaternions. Since the quaternions $\mathbf{q}$ and $-\mathbf{q}$ denote an identical rotation, we constrain the first quaternion in every sequence to the same hemisphere and interpolate subsequent quaternions to have the shortest distance along the unit hypersphere. We then split the interaction data into pairs of $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ windows to construct the samples for forecasting. We specify dataset-specific preprocessing details in Appendix 5.C.

Context Regimes. We evaluate on two context regimes: *random*, and *fixed-initial* (see Figure 5.6). In the *random* regime, context samples (observed-future pairs) are selected as a random subset of target samples, so the model is exposed to behaviors from any phase of the interaction lifecycle. Here we ensure that batches contain unique $\boldsymbol{t}_{\text{obs}}$ to prevent any single observed sequence from dominating the aggregation of representations over the context split. At evaluation, we take 50% of the batch as context. The *fixed-initial* regime investigates how models can learn from observing the initial dynamics of an interaction where certain gestures and patterns are more distinctive [1, Chap. 6]. Here we treat the first 20% of the entire interaction as context, treating the rest as target.

Conversation Groups as Meta-Learning Tasks? While our core idea of viewing groups as meta-learning tasks is grounded in social science literature (see Section section 5.5), does it help to improve empirical performance? Comparing the LL of non-meta-learning and meta-learning models in Table 5.2 by architecture—VED-MLP against NP and SP-MLP, and VED-GRU against SP-GRU—we find that accounting for group-specific dynamics through meta-learning yields improved performance. All best-in-family pairwise model differences are statistically significant (Wilcoxon signed rank test, $p < 10^{-4}$).

Comparing Within Meta-Learning Methods. While our SP-MLP models perform the best on LL in Table 5.2 (pairwise differences are significant), they fare the worst at estimating the mean (Appendix 5.A.2). On the other hand, the SP-GRU models estimate a better LL than the NP models with comparable errors in the mean forecast. The NP models attain the lowest errors in predicted means, but also achieve the worst LL. Why do the models achieving better LL also tend to predict worse means? Upon inspecting the metrics for individual features, we found that the models, especially the MLP variants, tend to improve LL by making the variance over constant features exceedingly small, often at the cost of errors in the means. Note that since the rotation in the data is in the horizontal plane, the qx and qy quaternion dimensions are zero throughout. We do not observe such model behavior in the synthetic data experiments, which do not involve constant features. Figure 5.7 visualizes forecasts for an example sequence from the Haggling dataset where a turn change has occurred just at the end of the observed window. Here, the SP-GRU model forecasts an interesting continuation to the turn. It anticipates that the buyer (middle) will interrupt the last observed speaker (right seller), before falling silent and looking from one seller to another, both of whom the model expects to then speak simultaneously (see Appendix 5.B for the full sequence). We believe that the forecast indicates that the model is capable of learning believable haggling turn dynamics from different turn continuations in the data. From the visualizations also we observe that the models seem to maximize LL at the cost of orientation errors; in the case of SP-MLP seemingly by predicting the majority orientation in the triadic setting. Also, the NP models forecast largely static futures. In contrast, while

Table 5.3: Mean (Std.) LL for the Ablation Experiments with the SP-uniform GRU Model. The reported mean and std. are over individual sequences in the test sets. Higher is better.

		MatchNMingle		Haggling	
		Random	**Fixed-Initial**	**Random**	**Fixed-Initial**
Full Model		87.0 (38.4)	85.5 (38.3)	79.9 (50.5)	78.6 (52.2)
Encoding Partner Behavior	no-pool	77.8 (31.2)	76.9 (31.0)	54.5 (75.5)	50.1 (97.5)
	pool-oT	82.3 (33.3)	81.0 (33.6)	66.9 (26.0)	66.8 (25.7)
No Deterministic Decoding	Shared Social Encoders	88.5 (40.7)	87.6 (39.6)	93.1 (39.3)	91.9 (40.4)
	Unshared Social Encoders	81.4 (38.1)	80.2 (37.8)	66.6 (24.0)	64.8 (23.4)

being more dynamic, the SP-GRU forecasts contain some smoothing. Overall, the SP-GRU models achieve the best trade-off between maximizing LL and forecasting plausible human behavior.

5

5.6.4 Ablations

Encoding Partner Behavior. Modeling the interaction from the perspective of each individual is a central idea in our approach. We investigate the influence of encoding partner behavior into individual representations $\boldsymbol{e}^i_{\text{ind}}$. We train the SP-uniform GRU variant in two configurations: *no-pool*, where we do not encode any partner behavior; and *pool-oT* where we pool over partner representations only at the last timestep (similar to [40]). Both configurations lead to worse LL and location errors (Table 5.3 and Appendix 5.A).

Deterministic Decoding and Social Encoder Sharing. We investigate the effect of the deterministic decoders by training the SP-uniform GRU model without them. We also investigate sharing a single social encoder between the Process Encoder and Process Decoder in Figure 5.2. Removing the decoders only improves log-likelihood if the encoders are shared, and at the cost of head orientation errors (Table 5.3 and Appendix 5.A).

5.7 Discussion

The setting of social conversations remains a uniquely challenging frontier for state-of-the-art low-level behavior forecasting. In the recent forecasting challenge involving dyadic interactions, none of the submitted methods could outperform the naive *zero-velocity* baseline [17, Sec. 5.5]. (The baseline propagates the last observed features into the future as if the person remained static.) Why is this? The predominant focus of researchers working on social human-motion prediction has been pedestrian trajectories [23] or actions such as *punching, kicking, gathering, chasing, etc.* [46, 47]. In contrast to such activities which involve pronounced movements, the postural adaptation for regulating conversations is far more subtle (also see the discussion in Appendix 5.E). At the same time, the social intelligence required to understand the underlying dynamics that drive a conversation

is comparatively more sophisticated than for an action such as a kick. We hope that the social-science considerations informing the design of SCF (joint probabilistic forecasting for all members) and the SP models (groups as meta-learning tasks) constitute a meaningful foundation for future research in this space to build upon. Note that for our task formulation, even the performance of our baseline models constitutes new results.

Cross-Discipline Impact and Ethical Considerations. While our work here is an *upstream* methodological contribution, the focus on human behavior entails ethical considerations for downstream applications. One such application involves assisting social scientists in developing predictive hypotheses for specific behaviors by examining model predictions. In these cases, such hypotheses must be verified in subsequent controlled experiments. With the continued targeted development of techniques for recording social behavior in the wild [81], evaluating forecasting models in varied interaction settings would also provide further insight. Another application involves helping conversational agents achieve smoother interactions. Here researchers should be careful that the ability to forecast does not result in nefarious manipulation of user behavior.

Acknowledgments

This research was partially funded by the Netherlands Organization for Scientific Research (NWO) under the MINGLE project number 639.022.606. Chirag would like to thank Amelia Villegas-Morcillo for her input and the innumerable discussions, and Tiffany Matej Hrkalovic for feedback on parts of the manuscript.

References

[1] A. Kendon. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Number 7 in Studies in Interactional Sociolinguistics. Cambridge University Press, Cambridge ; New York, 1990.

[2] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[3] D. Bohus and E. Horvitz. Models for multiparty engagement in open-world dialog. In *Proceedings of the SIGDIAL 2009 Conference on The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue - SIGDIAL '09*, pages 225–234, London, United Kingdom, 2009. Association for Computational Linguistics. doi: 10.3115/1708376.1708409.

[4] R. Ishii, S. Kumano, and K. Otsuka. Prediction of Next-Utterance Timing using Head Movement in Multi-Party Meetings. In *Proceedings of the 5th International Conference on Human Agent Interaction*, HAI '17, pages 181–187, New York, NY, USA, Oct. 2017. Association for Computing Machinery. doi: 10.1145/3125739.3125765.

[5] A. Keitel and M. M. Daum. The use of intonation for turn anticipation in observed conversations without visual signals as source of information. *Frontiers in psychology*, 6:108, 2015.

[6] S. Garrod and M. J. Pickering. The use of content and timing to predict turn transitions. *Frontiers in psychology*, 6:751, 2015.

[7] A. Rochet-Capellan and S. Fuchs. Take a breath and take the turn: how breathing meets turns in spontaneous dialogue. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658):20130399, 2014.

[8] M. Wlodarczak and M. Heldner. Respiratory turn-taking cues. In *INTERSPEECH*, 2016.

[9] D. Bohus and E. Horvitz. Managing Human-Robot Engagement with Forecasts and… um… Hesitations. *Proceedings of the 16th International Conference on Multimodal Interaction*, page 8, 2014.

[10] F. van Doorn. Rituals of Leaving: Predictive Modelling of Leaving Behaviour in Conversation. *Master of Science Thesis, Delft University of Technology*, 2018.

[11] L. Airale, D. Vaufreydaz, and X. Alameda-Pineda. SocialInteractionGAN: Multi-person Interaction Sequence Generation. *arXiv:2103.05916 [cs, stat]*, Mar. 2021.

[12] N. Sanghvi, R. Yonetani, and K. Kitani. Mgpi: A computational model of multiagent group perception and interaction. *arXiv preprint arXiv:1903.01537*, 2019.

[13] S. Bilakhia, S. Petridis, and M. Pantic. Audiovisual Detection of Behavioural Mimicry. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 123–128, Geneva, Switzerland, Sept. 2013. IEEE. doi: 10.1109/ACII.2013.27.

[14] C. C. S. Liem, M. Langer, A. Demetriou, et al. Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 197–253. Springer, 2018.

[15] E. Nilsen, D. Bowler, and J. Linnell. Exploratory and confirmatory research in the open science era. *Journal of Applied Ecology*, 57, Feb. 2020. doi: 10.1111/1365-2664.13571.

[16] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung. The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, 2018.

[17] C. Palmero, G. Barquero, J. C. J. Junior, et al. Chalearn lap challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 4–52. PMLR, 2022.

[18] C. Ahuja, S. Ma, L.-P. Morency, and Y. Sheikh. To React or not to React: End-to-End Visual Pose Forecasting for Personalized Avatar during Dyadic Conversations. *arXiv:1910.02181 [cs]*, Oct. 2019.

[19] M. Heldner and J. Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38

(4):555–568, Oct. 2010. doi: 10.1016/j.wocn.2010.08.002.

[20] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292, 1972. doi: 10.1037/h0033031.

[21] M. M. Moore. Nonverbal courtship patterns in women: Context and consequences. *Ethology and Sociobiology*, 6(4):237–247, Jan. 1985. doi: 10.1016/0162-3095(85)90016-0.

[22] N.-J. Moore, H. Mark III, and W. Don. Stacks. Nonverbal communication: Studies and applications. 2013.

[23] A. Rudenko, L. Palmieri, M. Herman, et al. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020.

[24] E. Goffman. *Behavior in Public Places: Notes on the Social Organization of Gatherings*. The Free Press, 1. paperback ed., 24. printing edition, 1966.

[25] A. Wang and A. Steinfeld. Group Split and Merge Prediction With 3D Convolutional Networks. *IEEE Robotics and Automation Letters*, 5(2):1923–1930, Apr. 2020. doi: 10.1109/LRA.2020.2969947.

[26] M. Mastrangeli, M. Schmidt, and L. Lacasa. The roundtable: An abstract model of conversation dynamics. *arXiv:1010.2943 [physics]*, Oct. 2010.

[27] H. Joo, T. Simon, M. Cikara, and Y. Sheikh. Towards Social Artificial Intelligence: Nonverbal Social Signal Prediction in a Triadic Interaction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10865–10875, Long Beach, CA, USA, June 2019. IEEE. doi: 10.1109/CVPR.2019.01113.

[28] S. Tan, D. M. J. Tax, and H. Hung. Multimodal Joint Head Orientation Estimation in Interacting Groups via Proxemics and Interaction Dynamics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–22, Mar. 2021. doi: 10.1145/3448122.

[29] N. T. V. Tuyen and O. Celiktutan. Context-aware human behaviour forecasting in dyadic interactions. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 88–106. PMLR, 2022.

[30] D. Helbing and P. Molnar. Social Force Model for Pedestrian Dynamics. *Physical Review E*, 51 (5):4282–4286, May 1995. doi: 10.1103/PhysRevE.51.4282.

[31] J. Wąs, B. Gudowski, and P. J. Matuszyk. Social Distances Model of Pedestrian Dynamics. In *Cellular Automata*, volume 4173, pages 492–501. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. doi: 10.1007/11861201_57.

[32] G. Antonini, M. Bierlaire, and M. Weber. Discrete Choice Models for Pedestrian Walking Behavior. *Transportation Research Part B: Methodological*, 40:667–687, Sept. 2006. doi: 10.1016/j.trb.2005.09.006.

[33] A. Treuille, S. Cooper, and Z. Popović. Continuum crowds. *ACM Transactions on Graphics / SIGGRAPH 2006*, 25(3):1160–1168, July 2006.

[34] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning Social Etiquette: Human

Trajectory Understanding In Crowded Scenes. In *Computer Vision – ECCV 2016*, volume 9912, pages 549–565. Springer International Publishing, Cham, 2016. doi: 10.1007/978-3-319-46484-8_33.

[35] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian Process Dynamical Models for Human Motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):283–298, Feb. 2008. doi: 10.1109/TPAMI.2007.1167.

[36] C. Tay and C. Laugier. Modelling Smooth Paths Using Gaussian Processes. In *Proc. of the Int. Conf. on Field and Service Robotics*, 2007.

[37] A. Patterson, A. Lakshmanan, and N. Hovakimyan. Intent-Aware Probabilistic Trajectory Estimation for Collision Prediction with Uncertainty Quantification. *arXiv:1904.02765 [cs, math]*, Apr. 2019.

[38] A. Alahi, K. Goel, V. Ramanathan, et al. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, Las Vegas, NV, USA, June 2016. IEEE. doi: 10.1109/CVPR.2016.110.

[39] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng. SR-LSTM: State Refinement for LSTM towards Pedestrian Trajectory Prediction. *arXiv:1903.02793 [cs]*, Mar. 2019.

[40] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. *arXiv:1803.10892 [cs]*, Mar. 2018.

[41] I. Hasan, F. Setti, T. Tsesmelis, et al. Forecasting People Trajectories and Head Poses by Jointly Reasoning on Tracklets and Vislets. *arXiv:1901.02000 [cs]*, Jan. 2019.

[42] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang. STGAT: Modeling Spatial-Temporal Interactions for Human Trajectory Prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6271–6280, Seoul, Korea (South), Oct. 2019. IEEE. doi: 10.1109/ICCV.2019.00637.

[43] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction. *arXiv:2002.11927 [cs]*, Feb. 2020.

[44] H. Zhao, J. Gao, T. Lan, et al. TNT: Target-driveN Trajectory Prediction. *arXiv:2008.08294 [cs]*, Aug. 2020.

[45] T. Gilles, S. Sabatini, D. Tsishkou, B. Stanciulescu, and F. Moutarde. THOMAS: Trajectory Heatmap Output with learned Multi-Agent Sampling. *arXiv:2110.06607 [cs]*, Jan. 2022.

[46] T. Yao, M. Wang, B. Ni, H. Wei, and X. Yang. Multiple Granularity Group Interaction Prediction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2246–2254, Salt Lake City, UT, June 2018. IEEE. doi: 10.1109/CVPR.2018.00239.

[47] V. Adeli, E. Adeli, I. Reid, J. C. Niebles, and H. Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters*, 5(4):6033–6040, 2020.

[48] Y.-W. Chao, J. Yang, B. Price, S. Cohen, and J. Deng. Forecasting Human Dynamics from Static

Images. *arXiv:1704.03432 [cs]*, Apr. 2017.

[49] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent Network Models for Human Dynamics. *arXiv:1508.00271 [cs]*, Sept. 2015.

[50] J. Walker, K. Marino, A. Gupta, and M. Hebert. The Pose Knows: Video Forecasting by Generating Pose Futures. *arXiv:1705.00053 [cs]*, Apr. 2017.

[51] I. Habibie, D. Holden, J. Schwarz, J. Yearsley, and T. Komura. A Recurrent Variational Autoencoder for Human Motion Synthesis. In *Procedings of the British Machine Vision Conference 2017*, page 119, London, UK, 2017. British Machine Vision Association. doi: 10.5244/C.31.119.

[52] D. Pavllo, D. Grangier, and M. Auli. QuaterNet: A Quaternion-based Recurrent Model for Human Motion. *arXiv:1805.06485 [cs]*, July 2018.

[53] M. Ranzato, A. Szlam, J. Bruna, et al. Video (language) modeling: A baseline for generative models of natural videos. *arXiv:1412.6604 [cs]*, Dec. 2014.

[54] J. Walker, A. Gupta, and M. Hebert. Dense Optical Flow Prediction from a Static Image. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2443–2451, Santiago, Chile, Dec. 2015. IEEE. doi: 10.1109/ICCV.2015.281.

[55] A. Dosovitskiy, P. Fischer, E. Ilg, et al. FlowNet: Learning Optical Flow with Convolutional Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, Santiago, Dec. 2015. IEEE. doi: 10.1109/ICCV.2015.316.

[56] J. Walker, A. Gupta, and M. Hebert. Patch to the Future: Unsupervised Visual Prediction. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3302–3309, Columbus, OH, USA, June 2014. IEEE. doi: 10.1109/CVPR.2014.416.

[57] C. Vondrick, H. Pirsiavash, and A. Torralba. Anticipating Visual Representations from Unlabeled Video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 98–106, Las Vegas, NV, USA, June 2016. IEEE. doi: 10.1109/CVPR.2016.18.

[58] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised Learning of Video Representations using LSTMs. *arXiv:1502.04681 [cs]*, Feb. 2015.

[59] A. Dosovitskiy and V. Koltun. Learning to Act by Predicting the Future. *arXiv:1611.01779 [cs]*, Nov. 2016.

[60] N. Ambady, F. J. Bernieri, and J. A. Richeson. Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In *Advances in experimental social psychology*, volume 32, pages 201–271. Elsevier, 2000.

[61] A. Vinciarelli, H. Salamin, and M. Pantic. Social Signal Processing: Understanding social interactions through nonverbal behavior analysis (PDF). *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, June 2009. doi: 10.1109/CVPRW.2009.5204290.

[62] A. Kalma. Gazing in triads: A powerful signal in floor apportionment. *British Journal of Social Psychology*, 31(1):21–39, Mar. 1992.

[63] S. C. Levinson and F. Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, June 2015. doi: 10.3389/fpsyg.2015.00731.

[64] E. Delaherche, M. Chetouani, A. Mahdhaoui, et al. Interpersonal Synchrony: A Survey of Evaluation Methods across Disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, July 2012. doi: 10.1109/T-AFFC.2012.12.

[65] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. Meta-Learning in Neural Networks: A Survey. *arXiv:2004.05439 [cs, stat]*, Nov. 2020.

[66] M. Garnelo, J. Schwarz, D. Rosenbaum, et al. Neural Processes. *arXiv:1807.01622 [cs, stat]*, 2018.

[67] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. 2014.

[68] K. Cho, B. van Merrienboer, C. Gulcehre, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv:1406.1078 [cs, stat]*, Sept. 2014.

[69] H. Kim, A. Mnih, J. Schwarz, et al. Attentive Neural Processes. *arXiv:1901.05761 [cs, stat]*, July 2019.

[70] G. Singh, J. Yoon, Y. Son, and S. Ahn. Sequential Neural Processes. *Advances in Neural Information Processing Systems*, 32, 2019. URL `http://arxiv.org/abs/1906.10264`.

[71] J. Yoon, G. Singh, and S. Ahn. Robustifying Sequential Neural Processes. In *International Conference on Machine Learning*, pages 10861–10870. PMLR, Nov. 2020.

[72] T. Willi, J. Masci, J. Schmidhuber, and C. Osendorfer. Recurrent Neural Processes. *arXiv:1906.05915 [cs, stat]*, Nov. 2019.

[73] S. Kumar. Spatiotemporal Modeling using Recurrent Neural Processes. page 43, 2019.

[74] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe. Analyzing Free-standing Conversational Groups: A Multimodal Approach. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 5–14. ACM Press, 2015. doi: 10.1145/2733373.2806238.

[75] L. Zhang and H. Hung. On Social Involvement in Mingling Scenarios: Detecting Associates of F-formations in Still Images. *IEEE Transactions on Affective Computing*, 2018.

[76] A. Kendall and R. Cipolla. Geometric Loss Functions for Camera Pose Regression with Deep Learning. *arXiv:1704.00390 [cs]*, May 2017.

[77] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention Is All You Need. *arXiv:1706.03762 [cs]*, June 2017.

[78] D. Ha and D. Eck. A Neural Representation of Sketch Drawings. *arXiv:1704.03477 [cs, stat]*, May 2017.

[79] S. R. Bowman, L. Vilnis, O. Vinyals, et al. Generating Sentences from a Continuous Space. *arXiv:1511.06349 [cs]*, May 2016.

[80] M. Vazquez, A. Steinfeld, and S. E. Hudson. Maintaining awareness of the focus of attention of a conversation: A robot-centric reinforcement learning approach. In *2016 25th IEEE International*

Symposium on Robot and Human Interactive Communication (RO-MAN), pages 36–43, New York, NY, USA, Aug. 2016. IEEE. doi: 10.1109/ROMAN.2016.7745088.

[81] C. Raman, S. Tan, and H. Hung. A modular approach for synchronized wireless multimodal multisensor data acquisition in highly dynamic social settings. *arXiv preprint arXiv:2008.03715*, 2020.

[82] C. Raman and H. Hung. Towards automatic estimation of conversation floors within f-formations. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 175–181. IEEE, 2019.

[83] T. A. Le, H. Kim, and M. Garnelo. Empirical Evaluation of Neural Process Objectives. In *NeurIPS workshop on Bayesian Deep Learning*, page 71, 2018.

[84] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, Jan. 2017.

[85] A. Paszke, S. Gross, F. Massa, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[86] W. Falcon et al. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3, 2019.

[87] R. Rienks, R. Poppe, and M. Poel. Speaker Prediction based on Head Orientations. In *Proceedings of the Fourteenth Annual Machine Learning Conference of Belgium and the Netherlands (Benelearn 2005)*, pages 73–79, 2005.

[88] M. Farenzena, A. Tavano, L. Bazzani, et al. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 30(2):115–127, May 2013. doi: 10.1111/j.1468-0394.2012.00622.x.

[89] S. Ba and J.-M. Odobez. Recognizing Visual Focus of Attention From Head Pose in Natural Meetings. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):16–33, Feb. 2009. doi: 10.1109/TSMCB.2008.927274.

Appendices

5.A Detailed Results

5.A.1 Forecasting Glancing Behavior: Quantitative Results

All models are evaluated under the *random* context regime and *no-pool* configuration. The sinusoids are interpreted to represent a horizontal head rotation between −90° and 90°. Figure 5.8 plots the LL and head orientation error per timestep in t_{fut}. In Figure 5.9 we plot the MAE in predicted and expected mean forecasts.

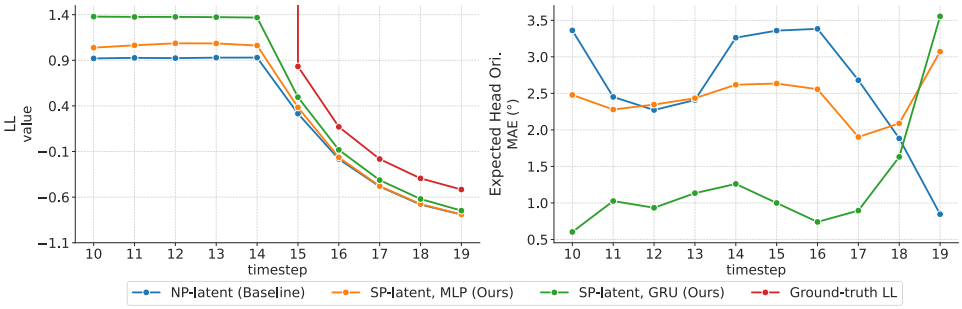


Figure 5.8: Mean Per Timestep Metrics over the Sequences in the Synthetic Glancing Dataset. We repeat Figure 5.5 here for completeness. Head orientation error is computed between the predicted and expected mean (mean of the two ground-truth futures). We observe that the SP-GRU model performs best, especially when the future is certain, learning both the best mean and std. over those timesteps.

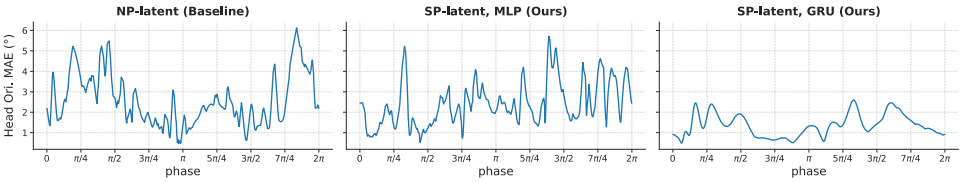


Figure 5.9: Error in forecast mean and expected mean orientation (mean of the two ground-truth futures) averaged over t_{fut} for every sequence in the Synthetic Glancing dataset. Each sequence is denoted by the phase of the sinusoid. The SP-GRU error plot is smoother with respect to small phase changes, with lower errors overall.

5.A.2 Errors in Predicted Means for Real-World Behavior Datasets

Tables 5.4 and 5.5 specify the error between the predicted mean forecast and ground-truth sequences in the test sets: mean-squared error (MSE) for the head and body keypoint locations; mean absolute error (MAE) in orientation in degrees; and speaking status accuracy. Note that we report the absolute error in rotation in 3D: while the ground-truth normals are constrained to the horizontal plane, we don't constrain our predicted quaternions. The metrics are computed by taking a mean over the participants, timestep dimensions of the tensors. The mean and std. are then reported over individual sequences.

Table 5.4: Mean (Std.) Errors in Predicted Means over Sequences in the MatchNMingle Test Sets. Lower is better. Underline indicates best measure within family.

(a) Random Context

Family	Model	Head Loc. MSE (px)	Body Loc. MSE (px)	Head Ori. MAE (°)	Body Ori. MAE (°)
VED [78, 79]	VED-MLP	131.12 (52.0)	110.58 (45.3)	74.84 (39.6)	97.55 (56.8)
	VED-GRU	30.00 (18.0)	24.80 (17.2)	22.62 (15.7)	24.72 (29.1)
NP [66, 69]	NP-latent	41.16 (21.4)	33.41 (21.1)	25.58 (21.1)	35.88 (47.7)
	NP-uniform	42.93 (20.9)	39.76 (21.0)	27.14 (21.4)	38.22 (48.0)
	ANP-dot	41.59 (20.4)	37.57 (19.5)	26.42 (20.6)	37.19 (47.6)
	ANP-mh	41.49 (20.7)	36.77 (19.5)	25.39 (20.7)	35.82 (48.0)
Ours (SP-MLP)	SP-latent	297.24 (92.3)	258.71 (87.8)	95.11 (41.3)	110.74 (51.6)
	SP-uniform	73.53 (34.2)	61.95 (36.2)	95.00 (41.3)	110.06 (50.8)
	ASP-dot	78.86 (24.8)	67.77 (21.7)	95.02 (41.3)	110.27 (51.1)
	ASP-mh	63.99 (22.4)	53.59 (22.2)	95.00 (41.3)	109.81 (50.6)
Ours (SP-GRU)	SP-latent	38.58 (20.9)	27.45 (19.9)	49.50 (44.1)	63.06 (57.4)
	SP-uniform	41.48 (22.2)	37.82 (18.9)	56.39 (47.8)	62.59 (54.3)
	ASP-dot	44.17 (21.4)	37.13 (19.7)	55.41 (47.1)	62.14 (54.5)
	ASP-mh	43.49 (21.3)	38.29 (19.8)	57.68 (47.3)	61.94 (53.7)

(b) Fixed-Initial Context

Family	Model	Head Loc. MSE (px)	Body Loc. MSE (px)	Head Ori. MAE (°)	Body Ori. MAE (°)
VED [78, 79]	VED-MLP	131.67 (52.6)	111.50 (46.0)	75.97 (38.7)	98.26 (55.6)
	VED-GRU	29.51 (16.5)	24.33 (16.1)	22.75 (15.9)	26.60 (32.0)
NP [66, 69]	NP-latent	40.82 (19.1)	32.81 (19.5)	25.58 (21.6)	38.97 (51.0)
	NP-uniform	45.22 (19.1)	40.60 (19.0)	28.34 (22.4)	42.65 (51.8)
	ANP-dot	44.67 (18.6)	40.03 (18.2)	29.08 (21.9)	44.44 (54.4)
	ANP-mh	42.75 (18.7)	37.56 (18.4)	26.95 (22.3)	42.20 (51.9)
Ours (SP-MLP)	SP-latent	296.36 (92.8)	259.46 (87.5)	94.75 (39.0)	108.62 (47.3)
	SP-uniform	81.61 (40.7)	64.44 (42.6)	94.68 (39.0)	108.26 (46.7)
	ASP-dot	92.03 (38.2)	78.97 (33.1)	94.69 (39.0)	108.36 (46.9)
	ASP-mh	66.22 (25.5)	53.04 (24.0)	94.67 (39.0)	108.14 (46.5)
Ours (SP-GRU)	SP-latent	38.31 (18.2)	26.79 (17.7)	51.78 (45.1)	65.38 (55.9)
	SP-uniform	42.75 (21.7)	42.18 (19.8)	57.79 (48.6)	64.44 (53.3)
	ASP-dot	54.42 (25.9)	44.88 (22.6)	56.12 (46.9)	65.28 (54.5)
	ASP-mh	56.62 (26.3)	47.78 (22.9)	58.90 (47.9)	64.46 (54.1)

Table 5.5: Mean (Std.) Errors in Predicted Means over Sequences in the Haggling Test Sets. Lower is better for all metrics except for speaking status accuracy. Underline indicates best measure within family.

(a) Random Context

Family	Model	Head Loc. MSE (cm)	Body Loc. MSE (cm)	Head Ori. MAE (°)	Body Ori. MAE (°)	Speaking Accuracy
VED [78, 79]	VED-MLP	42.04 (16.0)	41.53 (15.6)	24.70 (20.7)	19.02 (13.3)	0.636 (0.24)
	VED-GRU	0.79 (0.4)	0.75 (0.4)	1.55 (0.6)	1.06 (0.4)	0.989 (0.02)
NP [66, 69]	NP-latent	14.21 (6.5)	15.06 (6.1)	16.29 (13.8)	12.82 (13.7)	0.787 (0.23)
	NP-uniform	15.01 (7.3)	15.97 (7.2)	17.45 (18.3)	14.65 (20.0)	0.715 (0.24)
	ANP-dot	11.86 (5.4)	12.22 (5.5)	15.44 (13.3)	12.56 (18.0)	0.806 (0.23)
	ANP-mh	16.36 (7.4)	17.17 (7.2)	19.41 (20.4)	16.02 (22.1)	0.692 (0.21)
Ours (SP-MLP)	SP-latent	25.58 (10.1)	26.57 (9.0)	91.07 (23.9)	97.09 (22.5)	0.638 (0.08)
	SP-uniform	31.99 (8.2)	36.33 (7.3)	91.08 (23.9)	91.36 (23.9)	0.629 (0.18)
	ASP-dot	27.16 (7.7)	31.19 (7.1)	90.88 (23.9)	91.43 (23.8)	0.704 (0.19)
	ASP-mh	23.88 (7.8)	27.13 (7.7)	90.50 (23.9)	91.04 (24.1)	0.792 (0.24)
Ours (SP-GRU)	SP-latent	17.18 (6.5)	17.41 (6.2)	17.76 (15.8)	14.78 (20.7)	0.713 (0.23)
	SP-uniform	15.84 (5.5)	17.76 (7.5)	20.65 (19.9)	21.73 (29.5)	0.671 (0.22)
	ASP-dot	22.59 (8.7)	23.52 (10.2)	17.90 (11.3)	16.10 (19.3)	0.722 (0.24)
	ASP-mh	14.65 (5.8)	15.38 (6.1)	28.06 (24.5)	36.90 (37.9)	0.767 (0.23)

(b) Fixed-Initial Context

Family	Model	Head Loc. MSE (cm)	Body Loc. MSE (cm)	Head Ori. MAE (°)	Body Ori. MAE (°)	Speaking Accuracy
VED [78, 79]	VED-MLP	41.71 (16.2)	41.27 (15.8)	24.36 (19.8)	19.33 (13.4)	0.640 (0.25)
	VED-GRU	0.76 (0.4)	0.72 (0.3)	1.56 (0.6)	1.04 (0.3)	0.989 (0.02)
NP [66, 69]	NP-latent	13.85 (6.1)	14.71 (5.7)	16.22 (14.1)	12.69 (13.9)	0.774 (0.24)
	NP-uniform	15.01 (7.5)	15.95 (7.5)	17.26 (15.9)	14.68 (18.7)	0.701 (0.24)
	ANP-dot	12.83 (5.9)	13.26 (6.0)	16.19 (13.7)	13.56 (17.8)	0.717 (0.23)
	ANP-mh	16.68 (7.9)	17.43 (7.7)	19.78 (21.2)	15.57 (20.3)	0.682 (0.21)
Ours (SP-MLP)	SP-latent	25.27 (10.0)	26.33 (8.9)	91.14 (23.8)	97.09 (22.5)	0.640 (0.09)
	SP-uniform	32.93 (9.4)	37.16 (8.5)	91.15 (23.9)	91.36 (23.9)	0.633 (0.18)
	ASP-dot	27.94 (7.8)	31.83 (7.1)	90.93 (23.9)	91.43 (23.8)	0.628 (0.20)
	ASP-mh	24.07 (8.1)	27.35 (8.3)	90.53 (23.9)	91.07 (24.1)	0.770 (0.25)
Ours (SP-GRU)	SP-latent	16.66 (6.2)	17.17 (6.0)	17.67 (16.0)	14.64 (20.3)	0.705 (0.23)
	SP-uniform	16.53 (6.0)	18.20 (8.0)	20.74 (19.5)	21.31 (28.9)	0.674 (0.22)
	ASP-dot	23.91 (8.8)	25.34 (10.6)	19.11 (12.8)	17.36 (19.0)	0.635 (0.26)
	ASP-mh	16.87 (6.0)	16.96 (6.1)	28.90 (24.3)	37.23 (37.6)	0.705 (0.24)

The keypoint annotations for MnM are provided in image space from a top-down perspective, so the location errors in Table 5.4 are reported as the MSE in pixel locations. We do not consider speaking status cues for experiments with MnM (see Appendix 5.C.2).

5.A.3 Ablations

Table 5.6: Mean (Std.) Errors in Predicted Means for the Ablation Experiments with the SP-uniform GRU Model. The reported mean and std. are over sequences in the MatchNMingle Test Sets. Lower is better.

(a) Random Context

		Head Loc. MSE (px)	Body Loc. MSE (px)	Head Ori. MAE (°)	Body Ori. MAE (°)
Full Model		41.48 (22.2)	37.82 (18.9)	56.39 (47.8)	62.59 (54.3)
Encoding Partner Behavior	no-pool	36.25 (19.3)	30.88 (18.1)	47.28 (39.0)	62.09 (54.5)
	pool-oT	41.81 (19.7)	33.78 (17.9)	54.01 (45.3)	63.32 (54.8)
No Deterministic Decoding	Shared Social Encoders	41.84 (19.9)	30.99 (18.4)	44.59 (37.2)	72.02 (62.4)
	Unshared Social Encoders	37.25 (19.7)	36.13 (18.0)	62.81 (55.6)	56.15 (52.6)

(b) Fixed-Initial Context

		Head Loc. MSE (px)	Body Loc. MSE (px)	Head Ori. MAE (°)	Body Ori. MAE (°)
Full Model		42.75 (21.7)	42.18 (19.8)	57.79 (48.6)	64.44 (53.3)
Encoding Partner Behavior	no-pool	36.17 (17.4)	31.77 (16.5)	48.28 (39.6)	64.19 (53.3)
	pool-oT	41.91 (18.6)	34.17 (16.0)	54.95 (45.7)	65.20 (53.7)
No Deterministic Decoding	Shared Social Encoders	41.29 (18.2)	31.62 (16.9)	45.54 (38.0)	73.30 (60.9)
	Unshared Social Encoders	37.78 (18.5)	35.28 (16.3)	63.96 (56.1)	58.23 (53.0)

Table 5.7: Mean (Std.) Errors in Predicted Means for the Ablation Experiments with the SP-uniform GRU Model. The reported mean and std. are over sequences in the Haggling Test Sets. Lower is better for all except for speaking status accuracy.

(a) Random Context

		Head Loc. MSE (cm)	Body Loc. MSE (cm)	Head Ori. MAE (°)	Body Ori. MAE (°)	Speaking Accuracy
Full Model		15.84 (5.5)	17.76 (7.5)	20.65 (19.9)	21.73 (29.5)	0.671 (0.22)
Encoding Partner Behavior	no-pool	18.20 (6.7)	18.05 (7.7)	16.76 (12.8)	14.30 (20.9)	0.690 (0.21)
	pool-oT	17.02 (6.1)	19.18 (6.5)	23.71 (25.1)	17.80 (26.8)	0.738 (0.21)
No Deterministic Decoding	Shared Social Encoders	15.76 (7.2)	16.34 (6.6)	45.54 (44.6)	21.87 (25.0)	0.644 (0.22)
	Unshared Social Encoders	17.40 (6.9)	18.33 (6.7)	18.62 (14.7)	14.54 (20.2)	0.704 (0.23)

(b) Fixed-Initial Context

		Head Loc. MSE (cm)	Body Loc. MSE (cm)	Head Ori. MAE (°)	Body Ori. MAE (°)	Speaking Accuracy
Full Model		16.53 (6.0)	18.20 (8.0)	20.74 (19.5)	21.31 (28.9)	0.674 (0.22)
Encoding Partner Behavior	no-pool	18.64 (6.7)	18.45 (7.4)	16.85 (12.9)	14.29 (20.5)	0.687 (0.21)
	pool-oT	17.39 (6.2)	18.97 (6.4)	23.90 (24.6)	17.63 (25.6)	0.730 (0.21)
No Deterministic Decoding	Shared Social Encoders	16.93 (8.1)	17.15 (7.0)	45.49 (44.3)	21.83 (24.7)	0.637 (0.22)
	Unshared Social Encoders	18.54 (7.9)	19.18 (7.1)	18.68 (14.9)	14.44 (20.0)	0.700 (0.23)

5.B Qualitative Visualizations

5.B.1 Glancing Behavior

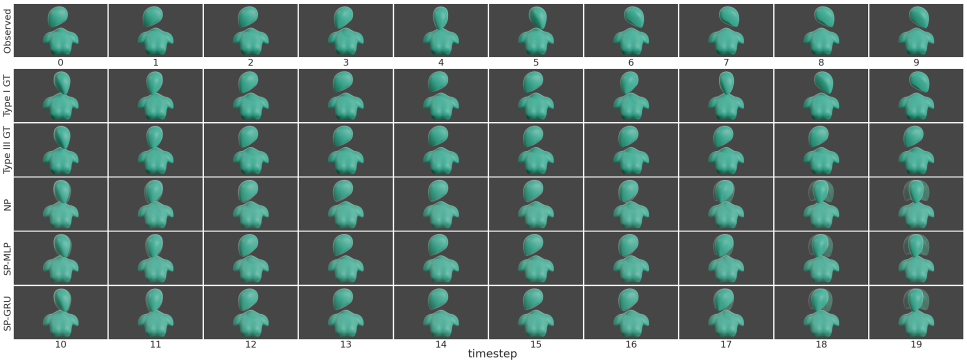


Figure 5.10: Forecasting Glancing Behavior for a Sequence in the Context Set. We visualize the same sinusoid within the context set as plotted in Figure 5.4 (phase = 4.2), here interpreted as a horizontal head rotation between −90° and 90°. The bottom three rows depict predictions, with the solid head denoting the mean, and the translucent heads the std. *GT* stands for *Ground-Truth*. The SP models learn better uncertainty estimates, especially over the timesteps where the future is certain (see timestep 11, for instance).

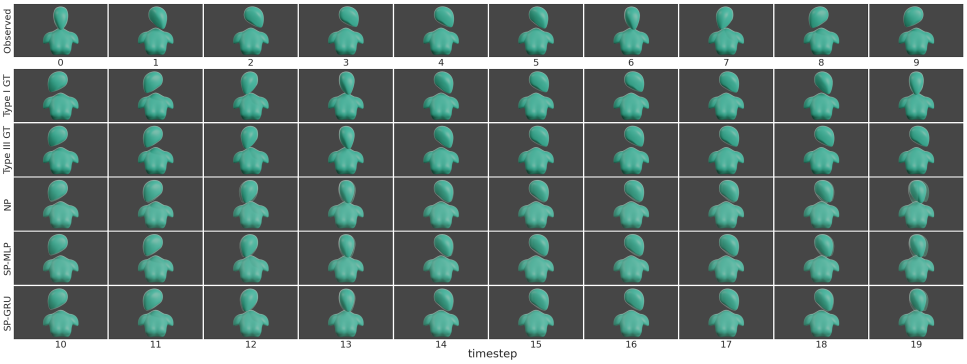


Figure 5.11: Forecasting Glancing Behavior for a Sequence Not in the Context Set. We visualize the same sinusoid not in the context set as plotted in Figure 5.4 (phase = 0.005). See the Figure 5.10 caption for details.

5.B.2 Real-World Behavior

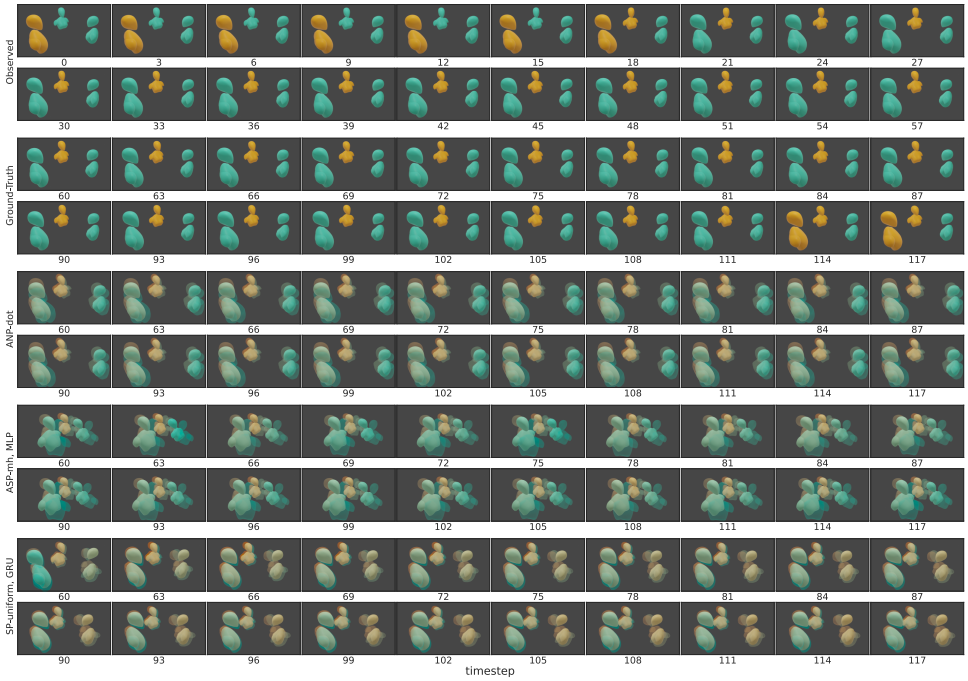


Figure 5.12: Forecasts for a Sequence from the Haggling Test Group *170221-b1-group3*. We visualize cues from real-world data using 3D models to preserve privacy. Similar to Figure 5.7, speakers are depicted in orange. The predicted speaking status mean is visualized as an interpolated shade between orange and blue. The translucent models in the forecasts denote the mean ± std. pose and speaking status. A speaker turn change occurs at around timestep 18 in the observed sequence. The buyer (on the right) looks at both sellers in turn mostly through gaze changes visible in the original video. This is barely registered in posture changes since both speakers are within the buyer's field of vision in this triadic setting (see Appendix 5.E for a discussion). The leaning motions of the new speaker, however, are captured in the postural shifts that continue into the ground-truth future. We observe that the NP forecasts are almost completely static. The SP-GRU forecasts are comparatively dynamic with lower uncertainties overall. The SP-MLP model seems to be learning an overall average orientation, forecasting all participants to be facing in the direction of the two sellers. Note that these pose changes are far more subtle than in the glancing behavior dataset, which is an important consideration for the domain practice of evaluating methods on synthesized behavior alone [12, 80].

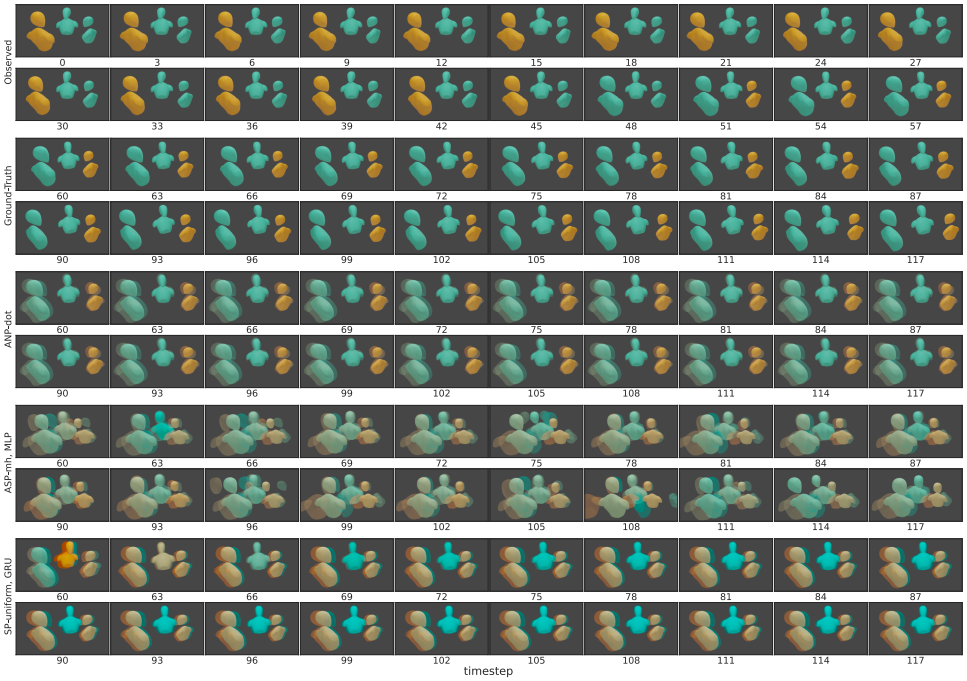


Figure 5.13: Forecasts for a Sequence from the Haggling Train Group *170224-a1-group1*. We see a similar pattern to the model forecasts as in Figure 5.12: NP forecasts are static, SP-GRU predicts more dynamic futures, while the SP-MLP forecasts average orientations. A turn change has occurred at the end of the observed window. We observe that the SP-GRU model forecasts an interesting continuation of the turn. It anticipates the buyer (middle) to quickly interject the last observed speaking seller, before falling silent and directing attention between the sellers, both of whom it expects to then speak simultaneously. While this is not the ground-truth future in this instance, we believe that the forecast still indicates that the model is capable of learning believable haggling turn dynamics from the overall training data. See the Figure 5.12 for details on the visualization setup.

5.C Additional Dataset Details

5.C.1 Synthesized Glancing Behavior Dataset

The set of pristine sinusoids representing *Type I* glances is computed by evaluating the sine function at the bounds of 19 equally spaced partitions of $[0, 3\pi + \phi)$, for phase values ϕ in $[0, 2\pi)$ with a step size of 0.001. More concretely, this is the set

$$g = \{r : r = \sin(x),\ x = n \times (3\pi + \phi)/19,\ n \in \{0, 1, \ldots 19\},$$
$$\phi = p \times 0.001,\ p \in \{0, 1, \ldots 6283\}\}, \quad (5.12)$$

which results in 6284 sequences. *Type III* glances are represented by identical sinusoids with clipped amplitudes for the last six timesteps, resulting in the final dataset of 12568

sequences. We train with batches of 100 sequences, using a randomly sampled 25 % of the batch as context. For evaluation, we fix 785 randomly sampled phase values as context. For each phase, samples corresponding to both types of glances are included in the context set, effectively using 25 % of all samples as context at evaluation.

5.C.2 Preprocessing the Real-World Behavior Datasets

For MnM, 2D keypoints are provided in image space (from a top-down camera perspective). For Haggling, the keypoints are in 3D space, and we use the nose keypoint to represent the head location, and the mid-point of the shoulders to represent the body location. We standardize the location features to have zero mean and unit variance, using the train statistics to standardize the test sets.

Haggling Preprocessing Details. Cue annotations are provided at 30 Hz for the Haggling Dataset. Motivated by the domain focus on the organization of turn-taking, we consider window lengths of 2 seconds supported by dataset statistics and literature. The Haggling dataset duration of contiguous speech follows a mean of 2.13 s (σ = 2.61 s), which is close to the mean measure of 1.68 s found in turn-taking analysis [19, 82]. We generate sliding windows with an overlap of 0.8, constraining the offset between $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ to a maximum of 5 s. This is to roughly restrict candidate future windows to those starting after two turn changes. In total, we obtain about 135K observed-future sequence pairs for training, and about 48K pairs for testing.

MatchNMingle Preprocessing Details. Cue annotations in MnM are provided at 1 Hz. The provided speaking status labels were annotated from video alone, and then manually smoothed by majority voting over 3 s windows. Consequently, these often do not match a person's pose behavior in the video for long sequences. We therefore deemed this data stream unsuitable for continuous sequence prediction and excluded it from our experiments. Assuming about 2 s per turn as before, and considering the 1 Hz annotation sample rate, we choose $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ to contain 4 timesteps or two turn durations each, with a maximum offset of 4 s as well. The keypoint annotations for every person are provided within the camera that best captures the individual, which can change over the duration of the interaction. For every group, we therefore first extract slices where the entire group is visible within the same camera for at least 20 s. We found 20 s to be a reasonable balance between not aggressively discarding groups, while still obtaining unique observed sequences for each slice (at least four). In total, we obtained about 74K observed-future sequence pairs for training, and about 52K pairs for testing.

5.D Implementation Details

5.D.1 Neural Architectures

The hyperparameters we chose resulted from light tuning through 5-fold cross-validation and showed improved performance for all models, but improved absolute performance might be obtained through more extensive tuning. The architecture hyperparameters were then kept fixed for the variants within each family for fair intra-family comparison. Table 5.8 specifies the network architecture hyperparameters for the real-world behavior dataset experiments. Note that for the MLP variants, the number of parameters is dependent on sequence length (timestep and feature dimensions of the tensors are collapsed into a single dimension; 60 timesteps for Haggling, 4 for MnM), so the final number of parameters vary across the datasets.

Table 5.8: Architecture Hyperparameters for real-world behavior dataset experiments (MnM / Haggling). For the meta-learning models, the number of parameters are reported for the simplest *-latent* variant.

Hyperparameter	VED-MLP	VED-GRU	NP	SP-MLP	SP-GRU
Sequence Encoder/Decoder					
Number of layers	2	1	2	2	1
Hidden dim	180	320	180/460	64	320
Partner Pooler $\psi(\boldsymbol{x}_j)$					
Number of MLP layers	—	—	—	2	2
MLP hidden dim	—	—	—	64	64
Output dim	—	—	—	32	32
***z* Encoder**					
Number of layers	2	2	2	2	2
Hidden dim	64	64	64	64	64
Representations					
$\boldsymbol{e}$, $\boldsymbol{r}$, $\boldsymbol{s}$, $\boldsymbol{z}$ dim	64	64	64	64	64
Multi-Head Attention					
Query/Key dim	—	—	32	32	32
Number of heads	—	—	8	8	8
Number of parameters					
MatchNMingle Dataset	254K	1.1M	274K	283K	3.0M
Haggling Dataset	711K	1.1M	2.8M	2.2M	3.0M

The non-meta-learning baselines retain the probabilistic attributes of our proposed Social Process models so that the only difference is the meta-learning aspect. We consequently adapt these baseline models from RNN based variational autoencoder architectures, first proposed for autoencoding sentences [79], and later refined for sketches [78]. The key difference is that rather than autoencoding the observed cues, we decode the future cues from the latent representations. Unlike [79], we are not working with discrete inputs, so the cues are fed directly into the sequence encoders without an embedding layer. For consistent comparison across models, we use unidirectional sequence encoders and decoders for the GRU variants and omit the Gaussian Mixture Model layer of [78]. This way, the encoding of

partner behavior is the only architectural difference in the backbone components between our proposed SP models and the VED baselines.

5.D.2 Training and Evaluation

The models are trained in the *random* context regime following the standard NP setting. We construct batches for training by bucketing samples such that all sequences in a batch share the same length of $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$. Note that since the MLP models are operationalized by collapsing the timestep and feature dimensions, the length of $\boldsymbol{t}_{\text{fut}}$ is fixed for these models across batches. However, since the recurrent models can handle sequences of different lengths, we allow for forecasting different length futures across batches, resulting in a few more training batches. Following the training practices suggested by Le et al. [83], we construct the context set at training as a random subset of the batch. Consequently, we further constrain samples in a batch to correspond to the same interacting group (see Section 5.5 for the underlying meta-learning intuition). For the same reason, we also ensure that a batch contains unique observed sequences so that a single observed sequence does not dominate the aggregation of representations over context. This is because a single observed sequence has multiple associated future sequences at different offsets, and could show up multiple times in a batch through random sampling if not handled explicitly.

We optimize the models using Adam [84]. For the NP and SP-MLP models we use a batch size of 128, an initial learning rate of $3 \cdot 10^{-5}$, a weight decay of $5 \cdot 10^{-4}$, and a dropout rate of 0.25. For the MLP-GRU models we use a batch size of 64, an initial learning rate of 10^{-5}, and a weight decay of 10^{-3}. The entire system was implemented using Pytorch [85] and Pytorch Lightning [86]. Every model was trained on a single NVIDIA GPU on an internal cluster depending on availability; one of Geforce GTX 970 (4 GB) or 1080 (8 GB), or Quadro P4000 (8 GB).

We validate the hyperparameters using 5-fold cross-validation, in the *random* context regime. At test, we use the same context sequences across models for a fair comparison. All testing was done with a batch size of 128 for consistency. The errors in mean are computed after destandardizing the location dimensions (orientation is already denoted by a unit quaternion, and therefore not standardized). The predicted std. deviations are scaled by the same value as the predicted means during destandardization.

5.E Distinguishing Forecasting in Focused and Unfocused Interactions: A Meta Discussion

Free-standing conversations are an example of what social scientists call *focused interactions*, said to arise when a "group of persons gather close together and openly cooperate to sustain a single focus of attention, typically by taking turns at talking" [24, p. 24]. On the other hand,

unfocused interactions occur when information is implicitly passed between individuals that happen to be in each other's presence by circumstance, such as pedestrians walking in proximity. One practical challenge of forecasting cues in focused interactions stems from the subtlety and sparsity of motion in recorded data. A common assumption is to use head pose as a proxy for gaze [12, 28, 74, 80, 87, 88]. In real-world data, however, attention shifts through changes in gaze are not always accompanied by similar head rotations [89, Fig. 5]. However, gaze is hard to record during group interactions in the wild with reasonable accuracy in a non-invasive manner. Even with the technology to do so (e.g. using onboard sensors on a social robot interaction partner), the question of whether recording faces is privacy-preserving is an ongoing discussion in the community. Moreover, intrusive sensing or non-human partners might also invalidate the naturalness of interaction behaviors (ecological validity). The consequence of not recording gaze is that in dyadic and triadic configurations where people are within each other's field of vision, the recorded movements (only from head and body) are even more subtle since attention shifts are predominantly achieved through gaze changes. This subtlety of motion in recorded data further distinguishes forecasting in conversations from the unfocused setting of pedestrian (or vehicle) trajectories. While some modeling techniques might be computationally applicable in both scenarios, the data stream in pedestrian trajectory settings (locations) can be comparatively more dynamic than the data streams in conversations (e.g. pose). It is important for researchers to be aware of such nuances while interpreting results for downstream applications (see Section 5.7).

6

Why Did This Model Forecast This Future? Information-Theoretic Saliency for Counterfactual Explanations of Probabilistic Regression Models

C. Raman, A. Nonnemaker, A. Villegas-Morcillo, H. Hung, and M. Loog. Why Did This Model Forecast This Future? Information-Theoretic Saliency for Counterfactual Explanations of Probabilistic Regression Models. *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, USA, 2023.

Abstract

We propose a post hoc saliency-based explanation framework for counterfactual reasoning in probabilistic multivariate time-series forecasting (regression) settings. Building upon Miller's framework of explanations derived from research in multiple social science disciplines, we establish a conceptual link between counterfactual reasoning and saliency-based explanation techniques. To address the lack of a principled notion of saliency, we leverage a unifying definition of information-theoretic saliency grounded in preattentive human visual cognition and extend it to forecasting settings. Specifically, we obtain a closed-form expression for commonly used density functions to identify which observed timesteps appear salient to an underlying model in making its probabilistic forecasts. We empirically validate our framework in a principled manner using synthetic data to establish ground-truth saliency that is unavailable for real-world data. Finally, using real-world data and forecasting models, we demonstrate how our framework can assist domain experts in forming new data-driven hypotheses about the causal relationships between features in the wild.

6.1 Introduction

As we go about our daily lives, engaging in conversations, walking down the street, or driving a car, we rely on our ability to anticipate the future actions and states of those around us [1, 2]. However, the numerous unknowns, such as hidden thoughts and intentions, make our predictions of the future inherently uncertain [2]. To reflect this uncertainty, several machine learning methods in such settings forecast a full distribution over plausible futures, rather than making a single point prediction [3, 4]. Identifying the factors that influence such a model's forecasts is particularly useful for domain experts seeking to understand the causal relationships guiding complex real-world behaviors, especially in situations where the future is uncertain. In this work, we introduce and address a novel research question toward counterfactual reasoning in multivariate probabilistic regression settings: how can we identify the observed timesteps that are salient for a model's probabilistic forecasts over a specific future window? Specifically, we introduce the first post hoc, model-agnostic, saliency-based explanation framework for *probabilistic* time-series forecasting.

We begin with a fundamental observation about human social cognition: we are averse to uncertainty and strive to minimize it [2]. Consider the scenario where a pedestrian is approaching you on the street. Initially, there is uncertainty about which direction each of you will take to avoid a collision. As one of you changes direction, the other observes and takes the opposite direction, ultimately avoiding a collision. Concretely, the thesis of this work is to formalize the following notion of saliency: the timestep that changes the uncertainty of a predicted future is salient toward predicting that future. For instance, in the aforementioned scenario, we posit that the moment when one pedestrian changes

direction is salient toward forecasting the future trajectories of the pedestrians.

Our notion of saliency is grounded in preattentive human cognition and related to the concept of surprisal or information associated with observations [5, 6]. Preattentive saliency captures what the brain subconsciously finds informative before conscious, or attentive, processing occurs. An unexpected or surprising observation is considered salient in this context. However, when applied to forecasting, the idea of surprisal or informativeness must be *linked to the future outcome*. Consequently, we propose that a timestep that alters an observer's certainty about the future is surprising, and therefore, salient. Crucially, our unifying 'bottom-up' perspective treats a forecasting model like a human observer, providing a principled definition of saliency that is not arbitrarily tied to task-specific error metrics. In contrast, the 'top-down' or task-specific notions of saliency common in post hoc explainable artificial intelligence (XAI) literature suffer from several drawbacks. Computed saliency maps may not measure the intended saliency, and even be independent of both the model and data generating process [7–9]. Moreover, what constitutes a *good* explanation is subject to the biases, intuition, or the visual assessment of the human observer [7, 10]; a phenomenon we refer to as the *interpretation being in the eye of the beholder*. Finally, as Barredo Arrieta et al. [11, Sec. 5.3] note, "there is absolutely no consistency behind what is known as saliency maps, salient masks, heatmaps, neuron activations, attribution, and other approaches alike."

To the best of our knowledge, no existing work addresses the specific task of obtaining post hoc model-agnostic explanations for probabilistic forecasts. Existing XAI methods for time-series data have predominantly focused on sequence classification, as we discuss in Section 6.2 and Appendix 6.A. For regression, instead of post hoc explainability, researchers have emphasized interpretability by design [11] or intrinsic interpretability [12], where interpretability stems from the simple structure of models or coefficients of predefined basis functions [13, 14]. Against this backdrop, we present the following key contributions:

- Conceptual Grounding: We establish the conceptual foundation for linking saliency-based explanations with counterfactual reasoning. We draw upon insights from Miller's [10] work on explanations in artificial intelligence, highlighting the contrastive nature of explanations (Section 6.3).
- Information-Theoretic Framework: We extend Loog's [5] framework of bottom-up preattentive saliency to the domain of probabilistic forecasting. Specifically, we introduce a novel expression of saliency based on the differential entropy of the predicted future distribution, providing a closed-form solution for commonly used density functions in the literature (Section 6.4).
- Empirical Validation: We empirically evaluate our framework using synthetic and real-world data. In the synthetic setting, we achieve full accuracy in retrieving salient timesteps with known ground truth saliency. In real-world scenarios without ground

truth saliency, we demonstrate the utility of our framework in explaining forecasts of social nonverbal behavior and vehicle trajectories, showcasing its effectiveness in complex and dynamic contexts (Section 6.5).

6.2 Related Work

XAI Techniques for Time-Series Data. The taxonomy commonly used for explainability methods categorizes techniques based on three criteria: (i) intrinsic or post hoc, (ii) model-specific or model-agnostic, and (iii) local or global [12]. In the context of time-series regression, existing techniques predominantly focus on non-probabilistic settings and fall into the category of intrinsic and model-specific approaches. These include: (i) incorporating inductive biases through internal basis functions [14] (also extended to the probabilistic setting [13]), (ii) utilizing self-attention mechanisms in the model [15], and (iii) adapting saliency maps from computer vision to measure the contribution of features to the final forecast [16, 17]. For a comprehensive review of XAI methods across domains and time-series tasks, please refer to Appendix 6.A.

Saliency-Based Explanations and Drawbacks. Saliency maps gained popularity as post hoc explanation tools for image classification [16, 18]. However, the lack of consistency in defining saliency has led to diverse interpretations, including occlusion sensitivity, gradient-based attribution heatmaps, and neuron activations [11, 12]. Nevertheless, these maps are typically computed by perturbing different parts of the input and observing the resulting change in the prediction error or output class. Several issues arise with the current use of saliency maps as explanations: (i) the feature-level manipulations used for saliency maps may distort the sample in ways that deviate from the real-world data manifold and destroy semantics [7–9]; (ii) given the arbitrary definitions, evaluating saliency maps becomes challenging and is subject to observer biases [12, Sec.10.3.2], which can lead to maps appearing correct even when they are insensitive to the model and data [7]; (iii) for forecasting, Pan et al.'s [17] notion of saliency based on the error between the point prediction and ground truth future is arbitrary and relies on ground truths unavialable during testing; and (iv) the saliency map is explicitly retrained for a single observed-future sequence, failing to capture salient patterns across similar observed sequences that result in divergent but plausible futures [17].

Model-Agnostic Techniques. The SHAP framework, which integrates ideas from Shapley Values, LIME, LRP, and DeepLIFT, has gained popularity as a model-agnostic approach [19]. However, adapting these techniques to time-series tasks poses several challenges. Firstly, the Shapley methods rely on functions with real-valued codomains, such as a regression function f_x [19, see Eq. 4, 8], while our focus is on probabilistic models that output the distribution $p_{Y|X}$ instead of some $y = f_x(\cdot)$ to handle future uncertainty. Adapting these

methods to deal with full predicted distributions is nontrivial. Similarly, gradient-based approaches compute gradients with respect to a single output instead of a full distribution. Secondly, these methods provide feature importance measures for a single output, whereas in time-series analysis, we are interested in identifying the importance of an observed timestep for an *entire future sequence*. That is, the joint consideration of the entire future sequence when computing input importance measures is challenging. As Pan et al. [17] note, in evaluating single-time predictions, these methods "ignore crucial temporal information and are insufficient for forecasting interpretation". Finally, similar to perturbation-based saliency methods, the sampling of features from arbitrary background samples in methods like Shapley/SHAP can lead to *Frankenstein Monster instances* [12, Sec. 9.5.3.3] that may not be valid samples on the data manifold. This undermines the semantics of the data, particularly in scenarios like motion trajectories, where randomly replacing features can result in physically impossible or glitchy motions.

6.3 Conceptual Grounding: Linking Saliency-Based Explanations to Counterfactual Reasoning

Given the challenges in XAI where speculations are often presented in the guise of explanations [20], we argue for grounding the concept of explanation within established frameworks of how humans define, generate, and present explanations. Turning to research in philosophy, psychology, and cognitive science, Miller [10] emphasized the importance of causality in explanatory questions. Drawing upon Pearl and Mackenzie's *Ladder of Causation* [21], he proposed the following categorization:

- Associative (*What?*): Reason about which unobserved events could have occurred given the observed events.
- Interventionist (*How?*): Simulate a change in the situation to see if the event still happens.
- Counterfactual (*Why?*): Simulate alternative causes to see whether the event still happens.

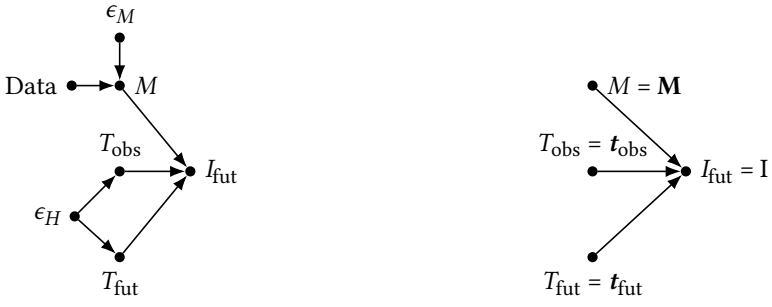
To apply Miller's framework in the context of forecasting, one needs to define the abstract notions of 'events' and 'causes'. Consider a model $\mathbf{M}$ that predicts features over a future window $\boldsymbol{t}_{\text{fut}}$ by observing features over a window $\boldsymbol{t}_{\text{obs}}$. We assert that the intrinsic interpretability methods involving inductive biases [13, 14] and attention mechanisms [15], fall under associative reasoning. These methods assess the (unobserved) importance of features over $\boldsymbol{t}_{\text{obs}}$ using model parameters or attention coefficients based on a single prediction from $\mathbf{M}$ (the 'event') for a fixed $\boldsymbol{t}_{\text{fut}}$ and single $\boldsymbol{t}_{\text{obs}}$. In contrast, we posit that the perturbation-based saliency methods can support counterfactual reasoning. They perturb different parts of the input over $\boldsymbol{t}_{\text{obs}}$ simulating alternative 'causes' from $\mathbf{M}$'s perspective,

and observe the effect on an error metric (the 'event'). However, the current application of these methods encounters issues outlined in Section 6.2.

To address the aforementioned challenges, we employ a unifying information-theoretic concept of bottom-up saliency grounded in preattentive human cognition [5, 6] as discussed in Section 6.1. Concretely, we propose the following implication that links this saliency to counterfactual reasoning:

observing the features at a timestep $t \in \boldsymbol{t}_{\text{obs}}$ results in a change in $\mathbf{M}$'s information about the future I_{fut} over the given $\boldsymbol{t}_{\text{fut}} \implies t$ is salient. (6.1)

Note that the antecedent (on the left of the implication) is a counterfactual statement. We formally express the implication using causal graphs [22] in Figure 6.1. The generic graph expresses relationships between the random variables prior to training the forecasting model M. The exogenous variable ϵ_M captures the randomness in the training process and modeling choices, including the distribution family for representing the forecasts. The exogenous variable ϵ_H captures the randomness in the human observer's choice of observed and future windows to examine the model. Our central idea is to evaluate the information in the model's predicted distribution denoted by I_{fut}. Specifically, we propose posing the following counterfactual question: *What information would* $\mathbf{M}$ *have about the future over a fixed* $\boldsymbol{t}_{\text{fut}}$ *if it observed the features over* $\boldsymbol{t}_{\text{obs}}$*?* The modified graph for evaluating this question is in Figure 6.1b. Once the model $\mathbf{M}$ has been trained and the windows $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ have been chosen, the effect of the exogenous variables on the variable I_{fut} disappears. This allows us to evaluate the change in the information about the future in response to *different* realizations of $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$, facilitating counterfactual analysis. Note that we assume the modified graph is already available, as our focus is on the explanation phase. While the procedure starting from training the model in the generic graph implicitly follows Pearl's *abduct-action-prediction* process [22, p. 207], estimating the distribution



(a) Generic Graph **(b)** Modified Graph

Figure 6.1: Causal Graphs for Explaining Forecasts

over the exogenous variables from the *abduction* step is conceptually not applicable in this setting.

Note that these graphs are not meant to describe relationships between random variables in the data for a specific hypothesis, as is typical in causal inference literature: for instance, the effect of [rotating toward the speaker] on [turn changes] in conversations. Rather, they describe the *process of a human generating contrastive explanations* for a given pretrained forecasting model M—irrespective of whether or not it is the true model—for some sequences in the data $(\boldsymbol{t}_{\text{obs}}, \boldsymbol{t}_{\text{fut}})$. Further, the notion of counterfactuals, as used within the context of contrastive explanations, is also distinct from that in causal inference. As Miller [10, Sec. 2.3] points out, "it is important to note that this is not the same counterfactual that one refers to when determining causality. For causality, the counterfactuals are hypothetical 'noncauses'...whereas in contrastive explanation, counterfactuals are hypothetical outcomes." Miller's point is that *why* explanations entail contrastive reasoning which involves comparing 'outcomes' in response to alternate 'causes'. In our work, this 'outcome' relates to the information in M's predicted distribution, the what-if question being "would the information in M's prediction for the window $\boldsymbol{t}_{\text{fut}}$ change if it had observed features over a different (contrastive) $\boldsymbol{t}_{\text{obs}}$?". Contrast this to associative reasoning which uses features from a single $\boldsymbol{t}_{\text{obs}}$ to generate the attribution map. A longer discussion is in Appendix 6.E.

Considering the information in forecasts in implication 6.1 links counterfactual reasoning to a more principled notion of saliency than has been used in XAI literature. Note that the implication entails that for the antecedent to be true t must be salient. However, knowing the antecedent is false is not sufficient to conclude that t is not salient, i.e. there can be other notions of saliency that make t salient. However, for less speculative evaluation, it is crucial that we use a unifying notion of saliency that is not arbitrarily defined based on task-specific error metrics or model gradients [11, 16–18]. Preattentive saliency, as we formalize in Section 6.4.1, is based on what is informative for the brain *before* conscious processing, making it more objective in nature.

Our framework addresses all the concerns associated with saliency-based approaches described in Section 6.2: (i) the counterfactuals in our framework are real observed features rather than random input perturbations, preserving the semantics of the real-world data; (ii) our use of information-theoretic preattentive saliency is principled and objective; (iii) our framework allows for saliency computation on unseen test data where the ground-truth future is unavailable, relying solely on the underlying model; and (iv) our approach considers the distribution over possible futures for a single input, capturing the structural predictive relationships between features across multiple samples. An additional advantage of our framework is that it does not require any training to compute the saliency and can be applied to any model that outputs a distribution over futures.

6.4 Methodology: Closed-Form Saliency for Probabilistic Forecasting

6.4.1 Preliminary: Information Theoretic Preattentive Saliency

Loog [5] developed a general closed-form expression for saliency based on computational visual perception that unifies different definitions of saliency encountered in the literature. The framework was illustrated on images and employed a surprisal-based operational definition of bottom-up attention. In this framework, an image is represented by a feature mapping function ϕ that relates each location in the image to a set of features. The saliency of a location x is determined by the information or surprise associated with its corresponding feature vector $\phi(x)$ compared to other feature vectors extracted from the same image. The saliency measure is defined as follows:

$$S(x) > S(x') \iff -\log p_\Phi(\phi(x)) > -\log p_\Phi(\phi(x')). \tag{6.2}$$

Here, p_Φ represents the probability density function over all feature vectors, while p_X captures any prior knowledge that influences the saliency of different image locations.

Contrary to approaches that determine saliency maps through an explicit data-driven density estimation [16, 18, 23–25], once the feature mapping ϕ is fixed, a closed-form expression for saliency can be obtained. The information content $-\log p_\Phi$ can be obtained from $\log p_X$ through a simple change of variables [26] from x to $\phi(x)$. The saliency $S(x)$ is then given by the expression:

$$-\log p_\Phi(\phi(x)) = -\log p_X(x) + \frac{1}{2}\log\det(J_\phi^t(x)J_\phi(x)), \tag{6.3}$$

where J_ϕ denotes the Jacobian matrix of ϕ, and $\_^t$ indicates matrix transposition. Since a monotonic transformation does not essentially alter the map, Loog [5] simplifies the saliency map definition to

$$S(x) \coloneqq \det(J_\phi^t(x)J_\phi(x)), \tag{6.4}$$

This formulation of saliency offers several advantages. It provides a principled and objective measure that captures the informativeness of features for human perception. Moreover, the saliency computation is purely local to an image, making it independent of previously observed data.

6.4.2 Defining ϕ in Terms of the Uncertainty over the Future Window $\boldsymbol{t}_{\text{fut}}$

Let $\boldsymbol{t}_{\text{obs}} \coloneqq [o1, o2, ..., oT]$ represent a window of consecutively increasing observed timesteps, and $\boldsymbol{t}_{\text{fut}} \coloneqq [f1, f2, ..., fT]$ denote an unobserved future time window, where $f1 > oT$. Con-

sider a set of n interacting agents, and let $\boldsymbol{X} \coloneqq [\boldsymbol{b}_t^i; t \in \boldsymbol{t}_{\text{obs}}]_{i=1}^n$ and $Y \coloneqq [\boldsymbol{b}_t^i; t \in \boldsymbol{t}_{\text{fut}}]_{i=1}^n$ represent their features over $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ respectively. Here, $\boldsymbol{b}_t^i$ captures multimodal features from agent i at time t. The forecasting task is to predict the density $p_{Y|X}$. Given a model that outputs $p_{Y|X}$, our task is to compute the saliency $S(\boldsymbol{t}_{\text{obs}})$ of an observed $\boldsymbol{t}_{\text{obs}}$ with respect to a fixed choice of $\boldsymbol{t}_{\text{fut}}$.

To extend Loog's [5] framework to forecasting settings, we need to choose an appropriate ϕ. We formalize the implication in Equation 6.1 and map $\boldsymbol{t}_{\text{obs}}$ to the differential entropy of the model's predicted future distribution over $\boldsymbol{t}_{\text{fut}}$. Specifically, we define $\phi : \boldsymbol{t}_{\text{obs}} \mapsto h(Y|X = \boldsymbol{X})$, where the conditional differential entropy of Y given $\{X = \boldsymbol{X}\}$ is defined as

$$h(Y|X = \boldsymbol{X}) \coloneqq -\int p_{Y|X}(\boldsymbol{Y}|\boldsymbol{X}) \log p_{Y|X}(\boldsymbol{Y}|\boldsymbol{X}) d\boldsymbol{Y}. \tag{6.5}$$

Our framework is summarized in Algorithm 1. Consider that a domain expert selects a specific $\boldsymbol{t}_{\text{fut}}$ corresponding to a high-order semantic behavior they wish to analyze. This could be a speaking-turn change [27, 28] an interaction termination [29, 30], or a synchronous behavior event [31]. Given an underlying forecasting model M and look-back period before $\boldsymbol{t}_{\text{fut}}$, we compute $h(Y|X = \boldsymbol{X})$ for different *observed multivariate features* $\boldsymbol{X}$ corresponding to different locations of a sliding $\boldsymbol{t}_{\text{obs}}$. The computed differential entropy values are then inserted into Equation 6.4 to obtain the saliency of different $\boldsymbol{t}_{\text{obs}}$ locations towards the future over the chosen $\boldsymbol{t}_{\text{fut}}$. In Appendix 6.B we discuss other favorable properties of differential entropy that make it a suitable choice as ϕ.

Explanation Using the Running Example. Within our running example from Section 6.1, $\boldsymbol{t}_{\text{fut}}$ corresponds to the two pedestrians passing each other while avoiding collision. In this example, let us assume M's training data contains examples of pedestrians passing others to both the left and the right. Consequently, for a $\boldsymbol{t}_{\text{obs}}$ containing the pedestrians approaching each other in a straight line, the predicted distribution $p_{Y|X}$ over $\boldsymbol{t}_{\text{fut}}$ encapsulates both possibilities of each pedestrian passing to the left as well as the right of the other. So the entropy $h(Y|X = \boldsymbol{X})$ is high for this $\boldsymbol{t}_{\text{obs}}$. Only once M is fed as input with the trajectories from the $\boldsymbol{t}_{\text{obs}}$ containing the pedestrians choosing one of the two

Algorithm 1 Temporal Saliency in Probabilistic Forecasting

Input: The probability density function $p_{Y|X}$, a fixed $\boldsymbol{t}_{\text{fut}}$ of interest, a sequence of m preceding observed windows $O = [\boldsymbol{t}_{\text{obs}}^1, \ldots, \boldsymbol{t}_{\text{obs}}^m]$, and the behavioral features $\boldsymbol{X}^j$ for every $\boldsymbol{t}_{\text{obs}}^j$

Output: The saliency map $S(O)$ over the observed windows

1: **for each** $\boldsymbol{t}_{\text{obs}}^j \in O$ **do**
2: Compute the feature mapping $\phi(\boldsymbol{t}_{\text{obs}}^j) \leftarrow h(Y|X = \boldsymbol{X}^j)$
3: **end for**
4: Compute saliency $S(\boldsymbol{t}_{\text{obs}}) \leftarrow \det(J_\phi^t(\boldsymbol{t}_{\text{obs}}) J_\phi(\boldsymbol{t}_{\text{obs}}))$

directions to pass, the predicted $p_{Y|X}$ is certain in terms of the pedestrians continuing along the chosen direction. (Note that in this case, we assume M has been trained by maximizing likelihood over the dataset containing only these two direction changes for avoiding collision.) Consequently, the entropy $h(Y|X = \boldsymbol{X})$ drops only once this moment of the pedestrians committing to a direction is seen by the model and would be considered salient for our algorithm.

6.4.3 Computing $h(Y|X = \boldsymbol{X})$

Typically, the density $p_{Y|X}$ is modeled as a multivariate Gaussian distribution [4, 32–34]. When the decoding of the future is non-autoregressive, the parameters of the distributions for all $t \in \boldsymbol{t}_{\text{fut}}$ are estimated at once, and the differential entropy has a closed-form expression, given by (see Cover and Thomas [35, Theorem 8.4.1])

$$h(Y|X = \boldsymbol{X}) = h(\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{K})) = \frac{1}{2}\log[(2\pi e)^d \det(\boldsymbol{K})]. \tag{6.6}$$

A common choice is to set $\boldsymbol{K}$ to be diagonal, i.e. the predicted distribution is factorized over agents and features. In this case, we can simply sum the log of the individual variances to obtain the feature mapping ϕ. Note that from Equation 6.6, for a multivariate Gaussian distribution, the differential entropy only depends on the covariance, or the *spread* of the distribution, aligning with the notion of differential entropy as a measure of total uncertainty. (See [35, Tab. 17.1; 36] for closed-form expressions for a large number of commonly employed probability density functions.)

In cases where probabilistic autoregressive decoders are used [4, 33, 37, 38], we do not have access to the full joint distribution $p_{Y_{f1},\ldots,Y_{fT}|X}$ for the timesteps in $\boldsymbol{t}_{\text{fut}}$. This is because inferring the density function $p_{Y|X}$ often involves sampling: a specific sample $\widehat{\boldsymbol{Y}}_t$ is taken from the predicted density at each $t \in \boldsymbol{t}_{\text{fut}}$, and passed back as input to the decoder for estimating the density at timestep $t+1$ [37, 38]. Therefore, the density at $t+1$ depends on the randomness introduced in sampling $\widehat{\boldsymbol{Y}}_t$. Figure 6.2 illustrates the concept for two timesteps. Here, a single forecast would only output the shaded red distribution for Y_2. In such cases, computing the joint entropy $h(Y_1, Y_2)$ directly is challenging in the absence of the full joint distribution p_{Y_1,Y_2}.

To address this, we have two options. The simpler option is to redefine our feature-mapping as $\phi : \boldsymbol{t}_{\text{obs}} \mapsto \sum_{t\in\boldsymbol{t}_{\text{fut}}} h(Y_t|\widehat{\boldsymbol{Y}}_{<t}, \boldsymbol{X})$, i.e. we approximate the total uncertainty over the predicted sequence by summing the differential entropies of the individual densities estimated at each timestep. Note that following the chain rule for differential entropy (see Cover and Thomas [35, Eq. 8.62]), the joint entropy can indeed be written as the sum of individual conditionals. However,

$$h(Y|X = \boldsymbol{X}) = \sum_{t\in\boldsymbol{t}_{\text{fut}}} h(Y_t|Y_{<t}, \boldsymbol{X}) \neq \sum_{t\in\boldsymbol{t}_{\text{fut}}} h(Y_t|\widehat{\boldsymbol{Y}}_{<t}, \boldsymbol{X}). \tag{6.7}$$

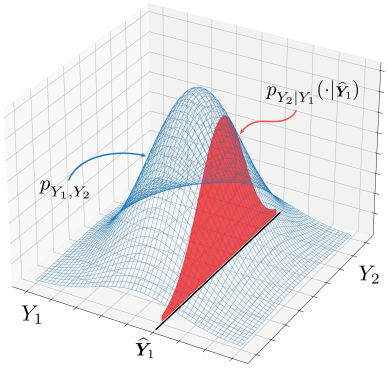


Figure 6.2: Illustrating predicted densities under greedy autoregressive decoding for two timesteps. For simplicity, we depict a joint Gaussian distribution and omit the conditioning on $\boldsymbol{X}$ everywhere.

And yet, training autoregressive decoders by maximizing likelihood actually assumes the inequality in Equation 6.7 to be approximately equal (see [39, Sec. 2; 40, Eq. 5]). The approximation relies on the observation that, for autoregressive decoding, the parameters of the predicted distribution for Y_t are computed as a deterministic function of the decoder hidden state. That is, Y_t is conditionally independent of $Y_{<t}$ given the hidden state of the decoder $\boldsymbol{s}_t$ at timestep t. The underlying assumption is that for a well-trained decoder, $\boldsymbol{s}_t$ encodes all relevant information from other timesteps to infer the distribution of Y_t. So at inference, despite being a function of the single sample $\widehat{\boldsymbol{Y}}_{t-1}$, the predicted distribution conditioned on $\boldsymbol{s}_t$ provides a reasonable estimate of the uncertainty in Y_t. This assumption allows us to again obtain a closed-form expression for the saliency map when each Y_t is modeled using a density function with a known closed-form expression for differential entropy [35, Tab. 17.1; 36]. For the common choice of modeling Y_t using a Gaussian mixture [37, 38], approximations that approach the true differential entropy can also be obtained efficiently [41–43] to directly compute the feature mapping ϕ.

The second option is to estimate $h(Y|X = \boldsymbol{X})$ using sampling or other non-parametric approaches when analytical expressions or computationally efficient approximations are not available [44–47]. These sampling-based methods provide approximations that converge to the true entropy, although they may be computationally more expensive than parametric methods. Overall, the choice of modeling the future density and the approach for estimating the differential entropy depends on the specific scenario and the available resources.

6.5 Experiments

The common evaluation of saliency-based explanations relies on qualitative visual assessment, which is subjective and prone to observer biases [7, 11, 12]. Meanwhile, establishing

a reliable ground truth for the salient relationship between the observed window $\boldsymbol{t}_{\text{obs}}$ and the future window $\boldsymbol{t}_{\text{fut}}$ is challenging in real-world data due to conflicting domain evidence on predictive relationships [28, 48]. Furthermore, fair validation of a *model agnostic, post hoc* method requires evaluating it independently of imperfections in the underlying forecasting model. To address these challenges we conduct two types of empirical evaluation: one using synthetic data to establish ground truth predictive saliency and *validate the framework*, and another to *demonstrate empirical utility* in real-world scenarios where perfect forecasts and ground truth saliency are unavailable.

No existing benchmarks or post hoc explanation frameworks exist for probabilistic time-series regression that meet the necessary requirements for a meaningful empirical comparison. Nevertheless, we provide results by adapting several explainability frameworks in our experiments. Specifically, we considered DeepSHAP and GradientSHAP [19], and IntegratedGradients and SmoothGrad [49]. It is important to note that we do not imply that these are fair comparisons; they are not (see Section 6.2). However, the comparisons are meant to characterize results from popular tools that practitioners are likely to use in the absence of our proposed framework. Implementation details for the following experiments and additional results for the real-world scenarios are in Appendices 6.C and 6.D, respectively.

6.5.1 Empirical Validation using Synthesized Ground Truth Saliency

Dataset. We simulate a group conversation that emulates real behavior patterns. Listeners typically focus on the speaker, while the speaker looks at different listeners [50]. Additionally, head gestures and gaze patterns predict the next speaker [51–54]. In our simplified simulation, the speaker rotates towards the center when speaking, and listeners nod to trigger a turn handover. We use real-valued quaternions to represent 3D head poses, commonly used for human motion and pose representation [4, 55, 56]. Following the notation in Section 6.4.2, $\boldsymbol{b} = [q_w, q_x, q_y, q_z, ss]$ where ss denotes binary speaking status. We simulate the turn changes to occur once clockwise and once anticlockwise. The ground truth salient timestep is when a listener initiates a head nod to trigger a turn handover, ensuring a certain future turn change. Figure 6.3 illustrates this mechanism. The code, dataset, and animated visualization are available in the Supplement.

Empirical Validation. To validate our framework in isolation, we assume a perfect forecasting model that predicts the true distribution over the possible future quaternion trajectories. The forecasting model focuses solely on low-level features and does not incorporate any high-order semantics of turn-taking. The saliency map generated by our framework, as shown in Figure 6.4a, accurately identifies the ground truth salient timesteps

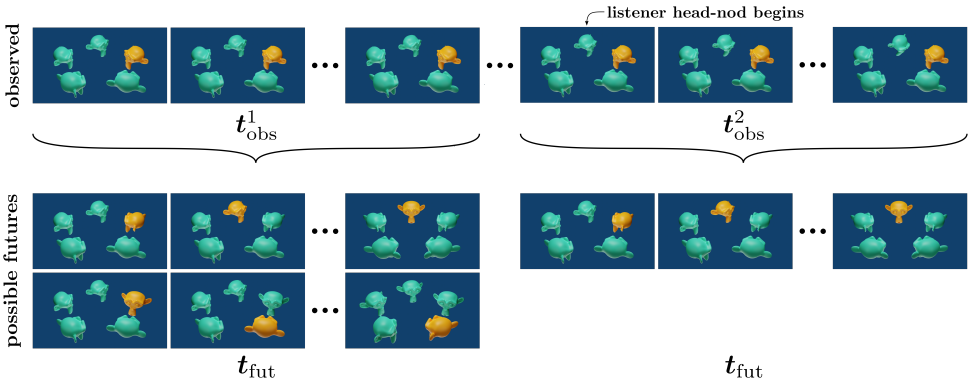
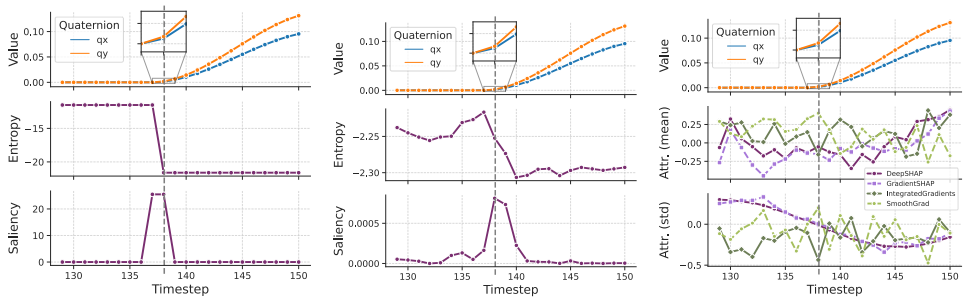


Figure 6.3: Illustrating the synthetic conversation dynamics dataset. Speakers are denoted in orange and listeners in green. For a fixed t_{fut} we depict two preceding t_{obs} windows. By construction, when observing a stable speaking turn over t^1_{obs}, two valid futures are possible over t_{fut}. These correspond to a turn handover to the immediate left or right of the current speaker. Over t^2_{obs}, when a listener nods to indicate the desire to take the floor, the future over t_{fut} becomes certain, corresponding to the listener successfully taking over the speaking turn. Here t^2_{obs} is consequently more salient than t^1_{obs} towards forecasting the turn change over t_{fut}. (Best viewed as video, see Supplement.)



(a) Ours (Perfect forecast) **(b)** Ours (Trained model) **(c)** Attributions (Trained model)

Figure 6.4: Computing Saliency. The top plots show the quaternion dimensions qx and qy for the listener that nods over t^2_{obs} in Figure 6.3. The gray dotted line indicates the true salient timestep 138 when the head nod begins, making the future over timesteps 183 – 228 (t_{fut}) certain. The rest of plots show the **(a)** entropy over future values of all participants (middle), and saliency map obtained using our framework (bottom), considering perfect forecasts; **(b)** entropy and saliency for the forecasts from a Social Process model; and **(c)** mean attributions across features per timestep from different explainability frameworks (DeepSHAP, GradientSHAP, IntegratedGradients, and SmoothGrad) for the predicted mean and std. of the same forecast from the Social Process model.

at frames 138 and 139 where the head nod begins. The saliency decreases once the nod is already in motion, indicating that it does not provide additional information about the future. This empirically validates our framework.

Introducing a Real Forecasting Model. We evaluate our framework using a real underlying forecasting model trained on synthetic data. We employ a *Social Process* model [4] for its ability to capture relative partner behavior and adapt to specific group dynamics. As shown in Figure 6.4b, our framework identifies the true salient timesteps with higher saliency values. Conversely, the attributions provided by other explainability frameworks in Figure 6.4c for the predicted mean and standard deviation of the same forecast fail to capture the salient predictive relationships in the data. This comparison underscores the effectiveness of our framework in capturing meaningful and interpretable saliency, even in conjunction with an imperfect forecasting model.

6.5.2 Empirical Evaluation in Real-World Scenarios

The study of group-leaving behavior has garnered interest in social psychology and the development of conversational agents [29, 30]. Recent approaches employ data-driven models to predict future non-verbal cues, capturing general predictive patterns in the data [4]. In this study, we demonstrate how our framework can assist domain experts in hypothesizing about the causal relationships between behavioral patterns and group leaving. We leverage the publicly available *MatchNMingle* dataset [57], which features natural interactions of 92 individuals during a cocktail party. We use an *Attentive Social Process* model [4] to forecast continuous head pose, body pose, and binary speaking status.

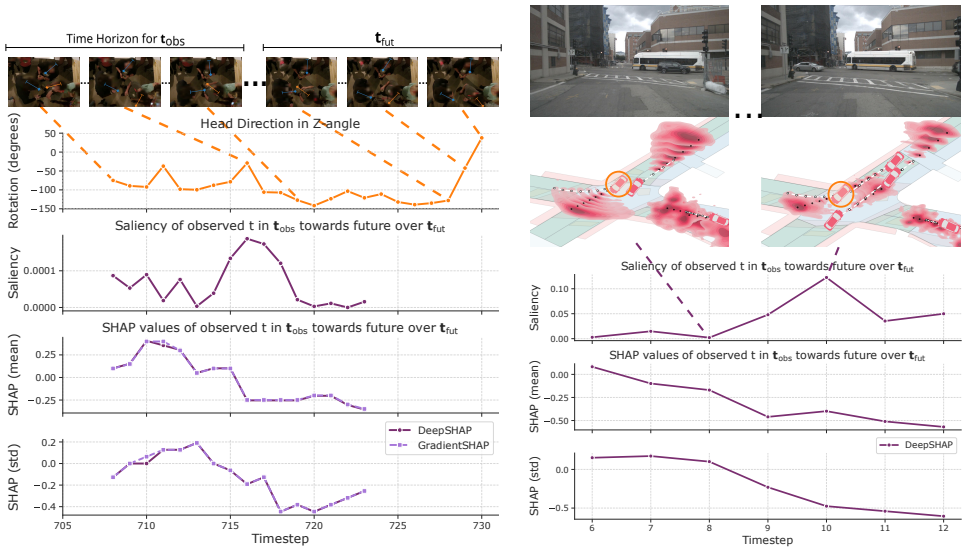
Through our analysis (see Figure 6.5a), we find that the salient timesteps in the model's forecasts correspond to instances when a person about to leave directs their gaze away from the shared group space (*o-space* [1]) by rotating their head. This observation leads to the following hypothesis:

> gazing away from the o-space of a conversing group is predictive of group leaving.

While this hypothesis aligns with established leave-taking patterns [1, 58] and the sweeping gaze behavior associated with seeking new interaction partners [59], it requires further validation through subsequent studies and rigorous statistical testing with the involvement of domain experts. Nonetheless, our experiment demonstrates how the framework can unveil data-driven insights into patterns that, in other cases, may have been overlooked by humans but captured by the forecasting model. By contrast, we do not observe any discernible intuitive patterns in the features associated with the trends in DeepSHAP and GradientSHAP values for the predicted mean and standard deviation.

Vehicle Trajectory Forecasting

The accurate forecasting of pedestrian and vehicle trajectories is crucial for safe and socially-aware autonomous navigation of vehicles [37, 60–62]. In this study, we utilize our framework to investigate vehicle dynamics in real driving scenarios. Specifically, we



(a) Group leaving behavior (*MatchNMingle* dataset) **(b)** Vehicle trajectory (*nuScenes* dataset)

Figure 6.5: (a) Analysis of the group leaving instance at 12:11 on *Day 1, Cam 12* in the *MatchNMingle* dataset. ***Row 1***: Video frames and overlaid arrows denoting the head orientation of participants. Orange indicates the person leaving the group over the t_{fut}. ***Row 2***: Head orientation of the leaver plotted as 2D horizontal rotation. ***Row 3***: Saliency map from running predictions from the Attentive Social Process model through our framework. The timesteps salient towards the model's forecasts correspond to the leaver making sweeping gazes away from the group. ***Rows 4-5***: mean DeepSHAP and GradientSHAP across features per timestep for the predicted mean and std. of the same forecast. **(b) Analysis of the vehicle turn making instance** on *Scene 3* in the *nuScenes* dataset. ***Row 1***: Video frames showing the bus and surrounding cars from the camera. ***Row 2***: Future predictions for the bus position (circled) from the Trajectron++ model (ground truth in white, predicted mean in black and variance in red). ***Row 3***: Saliency map from running predictions through our framework. The timesteps salient correspond to the model being more certain that the bus will make a turn. ***Rows 4-5***: mean DeepSHAP values across features per timestep for the predicted mean and std. of the same forecast. Best viewed as video (see Supplement).

leverage the *nuScenes* dataset, a multimodal dataset for autonomous driving [63], and the *Trajectron++* forecasting model [37].

Figure 6.5b illustrates our analysis of vehicle dynamics at an intersection. Notably, our framework identifies a salient timestep for the Trajectron++ model precisely when it becomes more confident that the bus will make a turn instead of continuing straight. This coincides with the model's increased certainty that the point-of-view vehicle will decelerate as a new vehicle enters the scene from the left. Although there are no relevant domain-specific theories in this case to interpret this saliency, these identified patterns align with expected driving behavior. In contrast, the DeepSHAP values fail to capture the model's change in certainty about the bus making the turn instead of continuing straight.

Moreover, we also do not identify any intuitive patterns in the predictions associated with the DeepSHAP trends. Thus, our framework serves as a valuable tool for sanity-checking model forecasts in real-world driving scenarios. It helps identify instances where the model's predictions align or misalign with established norms and expectations.

6.6 Conclusion

We have proposed a computational framework that provides counterfactual explanations of model forecasts based on a principled notion of bottom-up task-agnostic saliency. We derive a closed-form expression to compute this saliency for commonly used probability density functions to represent forecasts [4, 37, 38, 62]. To validate our framework, we conduct empirical experiments using a synthetic setup, enabling quantitative validation and mitigating observer biases associated with visual assessment of saliency maps. Additionally, we demonstrate the practical utility of our framework in two real-world scenarios involving the prediction of nonverbal social behavior and vehicle trajectories. By identifying salient timesteps towards a predicted future through counterfactual reasoning, our framework can support domain experts in formulating data-driven hypotheses regarding the predictive causal patterns involving the *features* present at those salient timesteps. These hypotheses can then be tested through subsequent controlled experiments, establishing a human-in-the-loop Explainable AI (XAI) methodology. For a more comprehensive discussion, please refer to Appendix 6.E.

6.7 Limitations and Potential Negative Societal Impact

While our framework provides a closed-form or efficient solution for most probability density functions, limitations arise when an analytic expression for differential entropy is unavailable. As discussed in Section 6.4.3, alternative approaches like sampling or nonparametric methods can be employed to approximate the entropy, albeit at an increased computational cost.

Our work here is an upstream methodological contribution. However, when applied downstream to human behavior or healthcare data, ethical considerations arise naturally. Here, care must be taken that such methods are not applied for gaining insights into behavior in a way that violates the privacy of people. Our framework enables domain experts to derive data-driven insights and hypotheses about predictive causal patterns. However, hypotheses should be rigorously tested, using controlled experiments and peer review, before being considered valid statements about human behavior. Collaboration among researchers, practitioners, and policymakers across disciplines is crucial to mitigate such societal risks and ensure ethical deployment of AI technologies.

Acknowledgments

The authors would like to thank Jesse Krijthe, Rickard Karlsson, David Tax, Yeshwanth Napolean, and Megha Khosla for the thoughtful discussions.

References

[1] A. Kendon. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Number 7 in Studies in Interactional Sociolinguistics. Cambridge University Press, Cambridge ; New York, 1990.

[2] O. FeldmanHall and A. Shenhav. Resolving uncertainty in a social world. *Nature human behaviour*, 3(5):426–435, 2019.

[3] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction. *arXiv:2002.11927 [cs]*, Feb. 2020.

[4] C. Raman, H. Hung, and M. Loog. Social Processes: Self-Supervised Meta-Learning over Conversational Groups for Forecasting Nonverbal Social Cues. *arXiv:2107.13576 [cs]*, July 2021.

[5] M. Loog. Information theoretic preattentive saliency: A closed-form solution. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1418–1424, Barcelona, Spain, Nov. 2011. IEEE. doi: 10.1109/ICCVW.2011.6130417.

[6] A. vanderHeijden. Perception for selection, selection for action, and action for perception, 1996.

[7] J. Adebayo, J. Gilmer, M. Muelly, et al. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[8] S. Lapuschkin, S. Wäldchen, A. Binder, et al. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nature Communications*, 10(1):1096, Dec. 2019. doi: 10.1038/s41467-019-08987-4.

[9] A. Atrey, K. Clary, and D. Jensen. Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep Reinforcement Learning, Feb. 2020.

[10] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, Feb. 2019. doi: 10.1016/j.artint.2018.07.007.

[11] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. doi: 10.1016/j.inffus.2019.12.012.

[12] C. Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[13] D. Rügamer, P. F. M. Baumann, T. Kneib, and T. Hothorn. Probabilistic Time Series Forecasts with Autoregressive Transformation Models. *arXiv:2110.08248 [cs]*, Feb. 2022.

[14] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv:1905.10437 [cs, stat]*, Feb. 2020.

[15] B. Lim, S. O. Arik, N. Loeff, and T. Pfister. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. *arXiv:1912.09363 [cs, stat]*, Sept. 2020.

[16] P. Dabkowski and Y. Gal. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30, 2017.

[17] Q. Pan, W. Hu, and N. Chen. Two Birds with One Stone: Series Saliency for Accurate and Interpretable Multivariate Time Series Forecasting. In *Twenty-Ninth International Joint Conference on Artificial Intelligence*, volume 3, pages 2884–2891, Aug. 2021. doi: 10.24963/ijcai.2021/397.

[18] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[19] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[20] Z. C. Lipton and J. Steinhardt. Troubling Trends in Machine Learning Scholarship, July 2018.

[21] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic books, 2018.

[22] J. Pearl. *Causality*. Cambridge university press, 2009.

[23] A. Torralba. Modeling global scene factors in attention. *JOSA A*, 20(7):1407–1418, 2003.

[24] N. D. Bruce. Features that draw visual attention: An information theoretic perspective. *Neurocomputing*, 65:125–133, 2005.

[25] L. Jiang, Z. Wang, M. Xu, and Z. Wang. Image saliency prediction in transformed domain: A deep complex neural network method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8521–8528, 2019.

[26] W. Boothby. An introduction to differential geometry and riemannian manifolds, 1975.

[27] A. Keitel, M. M. Daum, et al. The use of intonation for turn anticipation in observed conversations without visual signals as source of information. *Frontiers in psychology*, 6(108):265–273, 2015.

[28] S. C. Levinson and F. Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, June 2015. doi: 10.3389/fpsyg.2015.00731.

[29] D. Bohus and E. Horvitz. Managing Human-Robot Engagement with Forecasts and… um… Hesitations. *Proceedings of the 16th International Conference on Multimodal Interaction*, page 8, 2014.

[30] F. van Doorn. Rituals of Leaving: Predictive Modelling of Leaving Behaviour in Conversation. *Master of Science Thesis, Delft University of Technology*, 2018.

[31] S. Bilakhia, S. Petridis, and M. Pantic. Audiovisual Detection of Behavioural Mimicry. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 123–128, Geneva, Switzerland, Sept. 2013. IEEE. doi: 10.1109/ACII.2013.27.

[32] D. Salinas, M. Bohlke-Schneider, L. Callot, R. Medico, and J. Gasthaus. High-dimensional multivariate forecasting with low-rank Gaussian Copula Processes. In *Advances in Neural*

Information Processing Systems, volume 32. Curran Associates, Inc., 2019.

[33] D. Salinas, V. Flunkert, and J. Gasthaus. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. *arXiv:1704.04110 [cs, stat]*, Feb. 2019.

[34] S. S. Rangapuram, M. W. Seeger, J. Gasthaus, et al. Deep state space models for time series forecasting. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[35] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2006.

[36] A. V. Lazo and P. Rathie. On the entropy of continuous probability distributions (corresp.). *IEEE Transactions on Information Theory*, 24(1):120–122, 1978.

[37] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone. Trajectron++: Dynamically-Feasible Trajectory Forecasting With Heterogeneous Data. *arXiv:2001.03093 [cs]*, Jan. 2021.

[38] D. Ha and D. Eck. A Neural Representation of Sketch Drawings. *arXiv:1704.03477 [cs, stat]*, May 2017.

[39] I. Kulikov, S. Welleck, and K. Cho. Mode recovery in neural autoregressive sequence modeling. *arXiv preprint arXiv:2106.05459*, 2021.

[40] R. Dang-Nhu, G. Singh, P. Bielik, and M. Vechev. Adversarial attacks on probabilistic autoregressive forecasting models. In *International Conference on Machine Learning*, pages 2356–2365. PMLR, 2020.

[41] M. F. Huber, T. Bailey, H. Durrant-Whyte, and U. D. Hanebeck. On entropy approximation for Gaussian mixture random vectors. In *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 181–188, Seoul, Aug. 2008. IEEE. doi: 10.1109/MFI.2008.4648062.

[42] C. Zhang and Y. Luo. Approximating the differential entropy of gaussian mixtures. In *GLOBECOM 2017-2017 IEEE Global Communications Conference*, pages 1–6. IEEE, 2017.

[43] J. V. Michalowicz, J. M. Nichols, and F. Bucholtz. Calculation of Differential Entropy for a Mixed Gaussian Distribution. *Entropy*, 10(3):200–206, Sept. 2008. doi: 10.3390/entropy-e10030200.

[44] G. Ariel and Y. Louzoun. Estimating differential entropy using recursive copula splitting. *Entropy. An International and Interdisciplinary Journal of Entropy and Information Studies*, 22(2):236, 2020.

[45] B. J. Brewer. Computing entropies with nested sampling. *Entropy. An International and Interdisciplinary Journal of Entropy and Information Studies*, 19(8):422, 2017.

[46] J. Ajgl and M. Šimandl. Differential entropy estimation by particles. *IFAC Proceedings Volumes*, 44(1):11991–11996, Jan. 2011. doi: 10.3182/20110828-6-IT-1002.01404.

[47] J. Beirlant, E. J. Dudewicz, L. Györfi, E. C. Van der Meulen, et al. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1): 17–39, 1997.

[48] A. Kalma. Gazing in triads: A powerful signal in floor apportionment. *British Journal of Social*

Psychology, 31(1):21–39, Mar. 1992.

[49] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[50] H. Hung, D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 233–236, 2008.

[51] V. Petukhova and H. Bunt. Who's next? Speaker-selection mechanisms in multiparty dialogue. In *Workshop on the Semantics and Pragmatics of Dialogue*, 2009.

[52] I. De Kok and D. Heylen. Multimodal end-of-turn prediction in multi-party meetings. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, pages 91–98, 2009.

[53] U. Malik, J. Saunier, K. Funakoshi, and A. Pauchet. Who speaks next? Turn change and next speaker prediction in multimodal multiparty interaction. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 349–354. IEEE, 2020.

[54] R. Ishii, S. Kumano, and K. Otsuka. Prediction of Next-Utterance Timing using Head Movement in Multi-Party Meetings. In *Proceedings of the 5th International Conference on Human Agent Interaction*, HAI '17, pages 181–187, New York, NY, USA, Oct. 2017. Association for Computing Machinery. doi: 10.1145/3125739.3125765.

[55] G. Barquero, J. Núñez, S. Escalera, et al. Didn't see that coming: A survey on non-verbal social human behavior forecasting. *arXiv:2203.02480 [cs]*, Mar. 2022.

[56] D. Pavllo, D. Grangier, and M. Auli. QuaterNet: A Quaternion-based Recurrent Model for Human Motion. *arXiv:1805.06485 [cs]*, July 2018.

[57] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung. The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, 2018.

[58] M. L. Knapp, R. P. Hart, G. W. Friedrich, and G. M. Shulman. The rhetoric of goodbye: Verbal and nonverbal correlates of human leave-taking. *Communications Monographs*, 40(3):182–198, 1973.

[59] M. M. Moore. Nonverbal courtship patterns in women: Context and consequences. *Ethology and Sociobiology*, 6(4):237–247, Jan. 1985. doi: 10.1016/0162-3095(85)90016-0.

[60] T. Gilles, S. Sabatini, D. Tsishkou, B. Stanciulescu, and F. Moutarde. THOMAS: Trajectory Heatmap Output with learned Multi-Agent Sampling. *arXiv:2110.06607 [cs]*, Jan. 2022.

[61] S. Carrasco, D. F. Llorca, and M. Á. Sotelo. SCOUT: Socially-COnsistent and UndersTandable Graph Attention Network for Trajectory Prediction of Vehicles and VRUs. *arXiv:2102.06361 [cs]*, May 2021.

[62] A. Rudenko, L. Palmieri, M. Herman, et al. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020.

[63] H. Caesar, V. Bankiti, A. H. Lang, et al. nuScenes: A Multimodal Dataset for Autonomous Driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, Seattle, WA, USA, June 2020. IEEE.

[64] D. Baehrens, T. Schroeter, S. Harmeling, et al. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

[65] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[66] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[67] S. A. Bargal, A. Zunino, D. Kim, et al. Excitation Backprop for RNNs. *arXiv:1711.06778 [cs]*, Mar. 2018.

[68] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[69] F. Mujkanovic, V. Doskoč, M. Schirneck, P. Schäfer, and T. Friedrich. timeXplain – A Framework for Explaining the Predictions of Time Series Classifiers. *arXiv:2007.07606 [cs, stat]*, July 2020.

[70] H. Suresh, N. Hunt, A. Johnson, et al. Clinical intervention prediction and understanding using deep networks. *arXiv preprint arXiv:1705.08498*, 2017.

[71] A. A. Ismail, M. Gunady, H. C. Bravo, and S. Feizi. Benchmarking Deep Learning Interpretability in Time Series Predictions. *arXiv:2010.13924 [cs, stat]*, Oct. 2020.

[72] E. Choi, M. T. Bahadori, J. A. Kulas, et al. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. *arXiv:1608.05745 [cs]*, Aug. 2016.

[73] A. M. Alaa and M. van der Schaar. Attentive state-space modeling of disease progression. *Advances in neural information processing systems*, 32, 2019.

[74] M. Panja, U. Kumar, and T. Chakraborty. An Interpretable Probabilistic Autoregressive Neural Network Model for Time Series Forecasting. *arXiv:2204.09640 [cs, stat]*, Apr. 2022.

[75] L. Li, J. Yan, X. Yang, and Y. Jin. Learning Interpretable Deep State Space Model for Probabilistic Time Series Forecasting. *arXiv:2102.00397 [cs, stat]*, Jan. 2021.

[76] S. Joshi, S. Parbhoo, and F. Doshi-Velez. Pre-emptive learning-to-defer for sequential medical decision-making under uncertainty, Sept. 2021.

Appendices

6.A Broader Related Work: Explainable Methods for Time-Series Data Across Tasks and Domains

The larger focus of explainability techniques involving time-series data has been on the task of **classifying** time-series. The goal has been to estimate the relevance of each input feature at a given timestep towards each output class. Here, saliency approaches often overlap with techniques developed for image data and can be categorized into:

Gradient-Based Techniques. The broad approach involves evaluating the gradient of the output class with respect to the input [64]. Several variants have been proposed [19, 49, 65–67].

Perturbation-Based Techniques. The idea is to examine how the output changes in response to some perturbation of the input. Perturbations are implemented by either occluding contiguous regions of the input [68, 69]; performing an ablation of the features [70]; or randomly permuting features [12]. Ismail et al. [71] provide a benchmark of a subset of these techniques.

Attention-Based Techniques. These incorporate an attention mechanism into the model that is trained to attribute importance to different parts of the input sequence towards a prediction at each future timestep. Such techniques have been extensively utilized for healthcare data. Early methods applied a reverse-time attention [72], Later methods applied the attention to probabilistic state-space representations [73].

Some of these broad ideas have been applied to the **regression** setting to make interpretable forecasts of future time-series features. Lim et al. [15] leveraged self-attention layers for capturing long-term dependencies. Pan et al. [17] recently proposed computing saliency as a mixup strategy between series images and their perturbed version with a learnable mask for each sample. They view saliency in terms of minimizing the mean squared error between the predictions and ground-truths for a particular instance. Focusing on the univariate point-forecasting problem, Oreshkin et al. [14] proposed injecting inductive biases by computing the forecast as a combination of a trend and seasonality model. They argue that this decomposition makes the outputs more interpretable.

Developing explainable techniques for the probabilistic forecasting setting remains largely unexplored and subject to non-overlapping notions of explainability. Rügamer et al. [13] transform the forecast using predefined basis functions such as Bernstein polynomials. They relate interpretability to the coefficients of these basis functions (a notion similar to that of Oreshkin et al. [14]). Panja et al. [74] embed the classical linear ARIMA model into a non-linear autoregressive neural network for univariate probabilistic forecasting. As

before, the explainability here also stems from the 'white-box' nature of the linear ARIMA component. Li et al. [75] propose an automatic relevance determination network to identify useful exogenous variables (i.e. variables that can affect the forecast without being a part of the time-series data). To the best of our knowledge, saliency-based methods have not yet been considered within this setting.

6.B Favorable Properties of Differential Entropy

Differential entropy possesses favorable properties that make it a suitable choice as ϕ for computing the saliency map. First, the scale of the forecast density does not affect the resulting saliency map (see Cover and Thomas [35, Theorem 8.6.4]):

$$h(aY) = h(Y) + \log|a|, \text{for } a \neq 0, \text{and} \tag{6.8}$$

$$h(\boldsymbol{A}Y) = h(Y) + \log|\det(\boldsymbol{A})|, \text{when } \boldsymbol{A} \text{ is a square matrix.} \tag{6.9}$$

That is, scaling the distribution changes the differential entropy by only a constant factor. So the saliency map resulting from inserting the entropy into Equation 6.4 remains unaffected since the Jacobian term only depends on the relative change in entropy across different choices of $\boldsymbol{t}_{\text{obs}}$. Similarly, translating the predicted density leaves the saliency map unaffected (see Cover and Thomas [35, Theorem 8.6.3]):

$$h(Y + c) = h(Y). \tag{6.10}$$

6.C Implementation Details for Experiments

6.C.1 Other Explainability Methods

For DeepSHAP and GradientSHAP, we used the official implementation of SHAP: https://github.com/slundberg/shap. For IntegratedGradients and SmoothGrad, we used the Captum framework: https://captum.ai/. We reiterate that these are not fair comparisons, for the reasons we have discussed in Section 6.2. One crucial reason is that no existing method is designed to handle a predicted distribution. To apply them in this context, we need to compute attribution values for the predicted mean and standard deviation for every feature at every $t \in \boldsymbol{t}_{\text{fut}}$ in isolation. In contrast, by measuring differential entropy, our method jointly accounts for the parameters of the distribution and captures the information content of the distribution. Despite these limitations, we include these comparisons to provide readers with a contextual understanding of the results obtained from commonly used explainability tools.

Computational Efficiency. For further insight, we measured the execution time of our saliency method compared to DeepSHAP for both real-world scenarios (see Table 6.1).

Table 6.1: Comparing Computational Efficiency. We compare the practical execution time of our method to running DeepSHAP. For a reasonably fair characterization of DeepSHAP, we report execution time for computing the SHAP values associated with a single predicted parameter, the *mean* of the future distribution.

Scenario	Saliency		DeepSHAP	
	Forward Pass	Compute Saliency	Init	Compute Values
Group leaving (*MatchNMingle* dataset)	50 ms	2 ms	50 ms	13.5 min
Vehicle trajectory (*nuScenes* dataset)	2.5 s	33 ms	68 ms	26.3 min

The main takeaway is that our method is an order of magnitude faster at computing the saliency map compared to DeepSHAP. This is in part because DeepSHAP computes values for every feature at every timestep in the output independently. Even if this is parallelized, DeepSHAP requires multiple forward passes and gradient computations including samples from a background set to compute the reference values. The computation time scales with the size of the background set.

6.C.2 Empirical Validation using Synthetic Data

We model the future distribution using a Gaussian function for simplicity (setting std. to 10^{-10} for the single future), but a more complex distribution that predicts the appropriate change in variance would also work in practice. We now implement Algorithm 1 as follows. We identify a window where a turn change occurs in the data (frames 183-228) and denote this 45 frame window as the $\boldsymbol{t}_{\text{fut}}$ of interest. While we manually identify an interesting event for illustration, such a window could also correspond to an interesting prediction by a model. We generate a set of candidate $\boldsymbol{t}_{\text{obs}}$ by sliding a 30 frame window over a horizon of 100 frames prior to $\boldsymbol{t}_{\text{fut}}$, with a stride of 1 frame. For every observed $\boldsymbol{t}_{\text{obs}}$, we fit a Gaussian density to the quaternion and speaking status features of all participants over the futures that can occur during $\boldsymbol{t}_{\text{fut}}$. We then set the entropy of this Gaussian density as the feature ϕ for that $\boldsymbol{t}_{\text{obs}}$. For the experiments with a real forecasting model, we employ the `[recurrent, uniform attention]` variant of the Social Process family given its ability to capture dynamic movements [4].

6.C.3 Forecasting Group Leaving Behavior

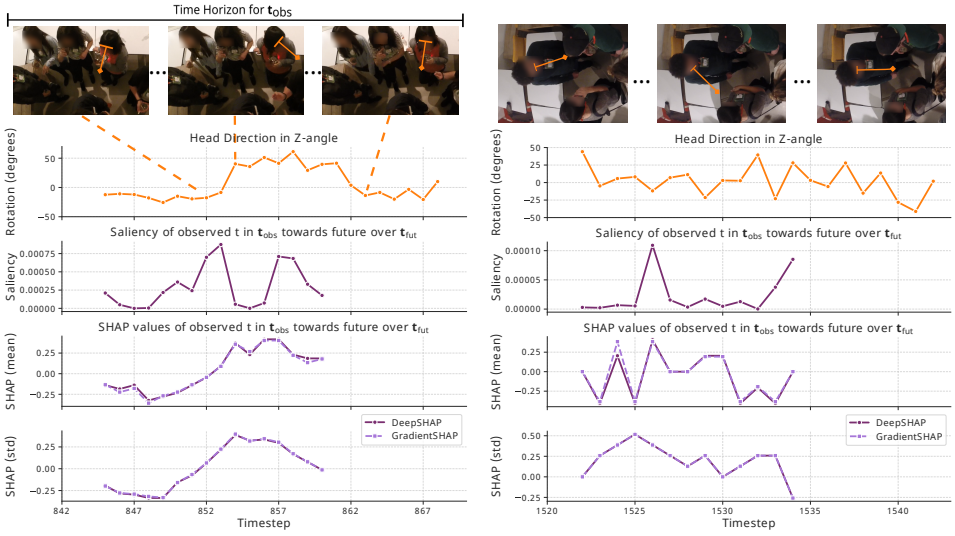
We used the pretrained model on the *MatchNMingle* dataset provided with the official implementation of Social Processes [4]: https://github.com/chiragraman/social-processes. Specifically, we employed the `[recurrent, dot-attention]` variant of the Attentive Social Process family (ASP-GRU-dot). We set $\boldsymbol{t}_{\text{fut}}$ to correspond to a 3 second window (3 frames in the data, which is at 1 Hz) containing an individual leaving a conversing group. We obtained forecasts from the model corresponding to a rolling 5 second $\boldsymbol{t}_{\text{obs}}$ within a 20 second preceding horizon. For computing DeepSHAP and GradientSHAP values, we used the entire observed time horizon as the background dataset. This ensures that the expected

values are computed within a reasonably similar context for a given t_{obs}.

6.C.4 Vehicle Trajectory Forecasting

We trained the Trajectron++ model [37] on the *nuScenes* dataset (mini version) [63] using the default command provided in the official implementation: https://github.com/StanfordASL/Trajectron-plus-plus. In particular, we used the `int_ee` model which incorporates the agent's system dynamics to produce dynamically-feasible trajectories. For analysis, we used a sequence from scene index 3 (`scene.name` 757). We set t_{fut} to correspond to a window of 6 timesteps (14-19) with a lookback horizon from timesteps 6 to 13. For computing DeepSHAP values we again used the entire observed horizon as the background set. The features in the observed sequence contained several `NaN` entries, which resulted in `NaN` DeepSHAP values. We ignored these when aggregating values per timestep.

6.D Additional Results



(a) 14:27 at *Day 1, Cam 12* **(b)** 25:42 at *Day 2, Cam 5*

Figure 6.6: Analysis of two sequences in the *MatchNMingle* dataset. ***Row 1***: Video frames and overlaid arrows denoting the head orientation of the participant of interest. ***Row 2***: Head orientation plotted as 2D horizontal rotation. ***Rows 3-5***: Saliency map from running predictions from the Attentive Social Process model through our framework, as well as the mean DeepSHAP and GradientSHAP values across features per timestep for the predicted mean and std. of the same forecast. Our saliency framework succeeds in identifying the salient timesteps in both cases. In **(a)**, the timestep in which the participant of interest is looking away is identified as salient towards predicting the other participant leaving the dyadic interaction. In **(b)**, the timestep in which the participant of interest suddenly stops actively participating in the conversation (not nodding or looking at the speakers) is identified as salient towards predicting their group leaving.

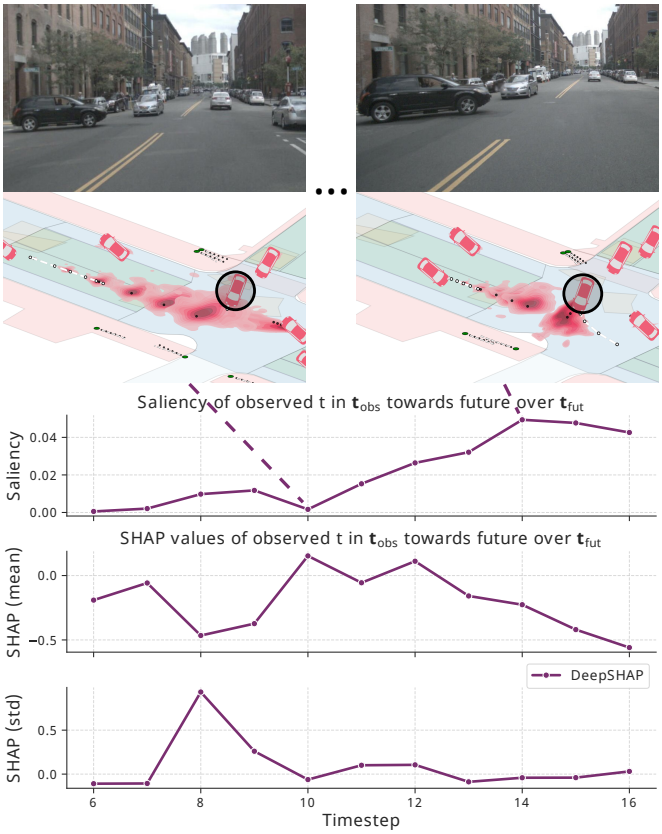


Figure 6.7: Analysis of a sequence on *Scene 0* in the *nuScenes* dataset. ***Row 1***: Video frames showing the black turning car and surrounding cars from the camera. ***Row 2***: Future predictions for the black car position (circled) from the Trajectron++ model (ground truth in white, predicted mean in black and variance in red). ***Rows 3-5***: Saliency map from running predictions through our framework, as well as the mean DeepSHAP values across features per timestep for the predicted mean and std. of the same forecast. Our framework reveals that the silver and point-of-view cars slowing is salient for the model in predicting that the black car completes the turn.

6.E Broader Discussion: Saliency & XAI with Domain Experts in the Loop

We begin this broader discussion by revisiting the different notions of saliency, to make the case for why our proposed framework is suitable for forecasting tasks. Rather than defining saliency in a top-down manner as a function of some task-specific error metric, we have started from a more fundamental conception of bottom-up, or task-agnostic, saliency. Loog's [5] original definition pertains to preattentive saliency, which captures what is perceived to be subconsciously informative before conscious (attentive) processing by the

brain. Here, a surprising or unexpected observation is salient. For instance, in a large white image with a single black pixel, the black pixel is salient. The direct application of this concept to time-series data would involve identifying surprising task-agnostic temporal events. For instance, imagine viewing a static landscape where a bird suddenly flies in. The entry of the bird into the scene is unexpected, and therefore salient.

When applied to forecasting tasks, however, this idea of surprisal (or unexpectedness or informativeness) that saliency represents needs to be tied to the future outcome. The saliency computed by most methods working on point-forecasting tasks deals with which past features are surprising given a specific realization of the future. While not explicitly stated by these works, we argue that this notion of saliency is related to the surprisal in $p_{X|Y}$ for some specific $\boldsymbol{Y}$. We therefore interpret these methods as being associative in nature within Miller's [10] categorization in Section 6.3. In contrast, our approach is counterfactual because we examine alternate future outcomes, while conceptualizing saliency more naturally defined in terms of the changes in the uncertainty in $p_{Y|X}$ in response to different realizations of observed sequences. However, rather than corresponding to random occlusions or perturbations of the input, the different realizations of X in our framework correspond to real features (or behaviors) preceding a future, which is more suitable to present to domain experts as candidate causes.

Loog's [5] unifying framework subsumes all forms of saliency, although identifying the appropriate ϕ for a specific domain is non-trivial. In this work we have established both theoretically and empirically how expressing ϕ in terms of the information about the future enables principled counterfactual reasoning in forecasting settings. Nevertheless, we reiterate that the salient timesteps retrieved by our framework ought to be treated as *candidate* causes until subsequently examined along with a domain expert. Our stance on human-in-the-loop XAI also aligns with research on saliency-based and general XAI in other domains [9, 76].

In principle, when it is possible to have access to the true $p_{Y|X}$, the salient timesteps identified by our framework reflect the *true* predictive structural relationships captured by the underlying model across the entire data. However, estimating this density analytically entails identifying the multiple futures in the data corresponding to every occurrence of the same observed features. In practice, subtle variations in behaviors and sensor measurement errors make it infeasible to estimate $p_{Y|X}$ analytically, so a model is trained to capture generalized patterns from the given data. In these cases, our framework identifies the sequences that *the model considers salient* for its forecasts *given the data.* Consequently, subsequent causal analysis of the features over the salient timesteps is crucial, especially in the healthcare and human behavior domains to avoid potential prejudices against certain behaviors, or worse, misdiagnoses of affective conditions.

III

PERCEPTION

Analyzing & Quantifying Social Phenomena

7

Where is the Conversation? Investigating the Existence of Multiple Conversation Floors within an F-formation

C. Raman and H. Hung. Towards Automatic Estimation of Conversation Floors within F-formations. *8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, Cambridge, United Kingdom, 2019, pp. 175-181. DOI: 10.1109/ACIIW.2019.8925065.

Abstract

The detection of free-standing conversing groups has received significant attention in recent years. In the absence of a formal definition, most studies operationalize the notion of a conversation group either through a spatial or a temporal lens. Spatially, the most commonly used representation is the F-formation, defined by social scientists as the configuration in which people arrange themselves to sustain an interaction. However, the use of this representation is often accompanied with the simplifying assumption that a single conversation occurs within an F-formation. Temporally, various categories have been used to organize conversational units; these include, among others, turn, topic, and floor. Some of these concepts are hard to define objectively by themselves. The present work constitutes an initial exploration into unifying these perspectives by primarily posing the question: can we use the observation of simultaneous speaker turns to infer whether multiple conversation floors exist within an F-formation? We motivate a metric for the existence of distinct conversation floors based on simultaneous speaker turns, and provide an analysis using this metric to characterize conversations across F-formations of varying cardinality. We contribute two key findings: firstly, at the average speaking turn duration of about two seconds for humans, there is evidence for the existence of multiple floors within an F-formation; and secondly, an increase in the cardinality of an F-formation correlates with a decrease in duration of simultaneous speaking turns.

Index Terms: *free-standing conversational groups, conversation floors, speaking turns*

7.1 Introduction

Imagine a social scenario like a mingling or networking event. Interactions in such a setting involve multiple dynamic conversations which are a medley of ever evolving topics and partners. And yet, humans can instinctively navigate the complexities of such encounters. How do we do this? We regulate our exchanges both spatially and temporally using implicit social norms or explicit behavioural signals [1]. Furthermore, these cues could be either verbal or non-verbal, expressed visually, vocally, or verbally through spoken language.

A deeper understanding of these group dynamics constitutes a natural objective towards the realisation of machines with social skills. For instance, consider a social robot approaching a group of people in a public space, or evaluating attendee experience at a conference poster session. In these and other cases, having an understanding of the dynamics, and where channels of social influence lie, would enable the artificial agent to develop increasingly sophisticated policies for interaction or inference. Conversation groups have been of importance in the application domains of social robotics [2–5], activity recognition [6, 7], social surveillance [8–10], and social signal processing [11, 12].

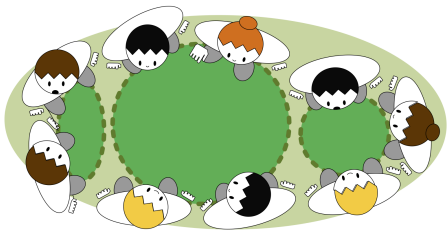


Figure 7.1: Depiction of a single F-formation with multiple conversation floors. The darker green regions within dotted lines represent distinct simultaneous conversation floors. Most works representing a conversing group as an F-formation make the simplifying assumption that a single conversation occurs within an F-formation with a joint focus of attention for all members.

Fundamental to the study of such conversations is defining the notion of a free-standing conversational group (FCG). While it is easier to objectively conceptualize an FCG in spatial terms in a scene of multiple interacting groups, delineating the boundary of conversations poses a greater technical challenge. We could think of separating conversations on the basis of topics, but this is challenging if audio data is unavailable due to privacy concerns. We could operationalize a conversation as a set of participating members, but this membership is challenging to infer visually for non-speaking participants. This often leads to the simplifying assumption in some literature that the focus of an FCG is a single conversation. As we illustrate in Figure 7.1, and discuss in the following sections, this may not always be the case.

In the present work, we dive beyond the geometric bounds of an FCG to gain a deeper understanding of the conversations occuring within it. In this initial approach, we focus specifically on speaking participants as the most decisive indicator of the existence of a conversation. Concretely, we pose the following broad research questions:

RQ 1. Can we use observed speaker turns to infer the conversation floors within an F-formation?

RQ 2. How does the cardinality of an F-formation affect the conversation floors developed within it?

The ground truth for speaker turns in this work comes from manual annotations of video data, mimicking use-cases where audio data might be unavailable due to privacy concerns. Concretely, our contributions are as follows: conceptually, we provide an indicator of distinct conversation floors that uses speaking turns alone, and situate this indicator in schisming literature [13–15]; analytically, we provide evidence that multiple conversation floors exist within an F-formation, and show that the cardinality of an F-formation correlates negatively with turn duration of simultaneous speakers.

The rest of this paper is organized as follows. We describe some of the spatial and

temporal perspectives used to study FCGs in Section 7.2. In Section 7.3 we provide a review of literature involving the use of these spatial or temporal notions, motivating the need to consider both of these aspects in unison. In Section 7.4, we propose an operationalization of an indicator of distinct conversation floors, building upon the concepts of conversation schisming. The dataset we use and the experiments performed for answering the research questions are described in Section 7.5 and Section 7.6 respectively. Finally, Section 7.7 summarizes our findings and concludes the paper.

7.2 Background

Spatial Factors. One of the most common proxemic notions to describe an FCG is Adam Kendon's *Facing Formation*, or *F-formation*, originally defined as:

> An F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access. [16, p. 210]

Kendon argues that activity is always located, and denotes the space in front of a person that is used for the activity as the person's *transactional segment*. When two or more people come together to perform some activity, they are liable to arrange themselves such that their individual transactional segments overlap to create a joint transactional space. This joint space between the interactants is called an *o-space*. As we discuss in the next section, many computational works involving the automatic detection of FCGs from video focus on the detection of F-formations, often assuming that the transaction involves a single conversation.

Temporal Factors. The conversation of focus in an FCG, however, is dynamic in nature. If conversations change over time, what are the temporal units that describe their underlying structure? Some of the terms used in early literature to organize conversational units include *turn*, *topic*, *gap*, and *floor*. Edelsky provides an excellent review of these concepts in [17], stating that most of these units were defined on the basis of some technical or mechanical structure such as signals of speakers or auditors, ignoring the intention of the participant. Using inferred participants' meanings rather than technical definitions, Edelsky defines turns and floors as follows:

> The floor is defined as the acknowledged what's-going-on within a psychological time/space. What's going on can be the development of a topic or a function (teasing, soliciting a response, etc.) or an interaction of the two. It can be developed or controlled by one person at a time or by several simultaneously or in quick succession. [17, p. 405]

7.3 Related Work

Detecting Conversational Groups. In most works, a conversational group is operationalized as an F-formation. Early work on the task of detecting FCGs in video data developed concurrently from two perspectives: those that estimate the location of the o-space using a Hough-voting strategy [8, 18]; or those that view an F-formation as a set with individuals being assigned exclusive membership [12, 19]. There has also been considerable work focused on incorporating temporal information for the same task of detecting conversational groups [7, 20–22]. Notably, these approaches utilise the head pose as a proxy for Visual Focus of Attention (VFoA) [9] in addition to the body pose to model F-formation membership, and assume a single conversation within an F-formation. The assumption that members in a group have a single joint focus of attention is seen in other works as well. Hung et al. [23] model a single joint focus of visual attention of participants to estimate dominance in groups. Vazquez et al. [4] also assume a single conversation within an F-formation while developing a policy for a robot to be aware of a single focus of attention of the conversation.

Estimating involvement. In a conversation, the floor is typically held by a single participant at a time [13]. What then characterizes the silent participants in a conversation group? The following works demonstrate that the task of estimating participant involvement is subjective in nature, and that gaze behaviour and turn-taking patterns can be informative.

Zhang and Hung [24, 25] study the task of detecting associates of an F-formation; members that are attached to an F-formation but do not have full status [16]. They argue that the labeling of conversation groups is not an objective task. Collecting multiple annotations of perceived associates, they demonstrate how detecting them can improve initial estimates of full-members of an F-formation. Oertel et al. [26] characterize silent participants into multiple categories (attentive listener, side participant, bystander) from audiovisual cues. Oertel and Salvi [27] also show that it is possible to estimate individual engagement and group involvement in a multiparty corpus by analysing the participants' eye-gaze patterns. Bohus and Horvitz [28] propose a self-supervised method for forecasting disengagement with an interactive robot using a conservative heuristic. The heuristic is constructed by leveraging features that capture how close the participant is, whether a participant is stationary or moving, and whether a participant is attending to the robot.

Some works also used turn-taking features to estimate some notion of involvement. Pentland et al. [29] measured engagement by the z-scored influence each person has on the other's turn-taking for a pair of participants. Hung and Gatica-Perez [30] found that the pause duration between an individual's turns, aggregated at group level, is highly predictive of cohesion in small group meetings.

Schisming. In a conversation with at least four participants, the conversation sometimes splits up into two or more conversations. This transformation is referred to as a *schism* [13] or *schisming*. One of the earliest allusions to the phenomenon of schisming based on anecdotal evidence occurs in the work of Goffman, who suggested that a gathering of two participants *exhausts* an encounter and forms a *fully-focused gathering* [31, p. 91]. With more than two participants, there may be persons officially present in the situation who are not themselves so engaged. These *bystanders* change the gathering into a *partly-focused* one. If more than three persons are present, there may be more than one encounter carried on in the same situation, resulting in a *multifocused* gathering.

In subsequent work, Sacks et al. [13] and Goodwin [14] both indicated that the co-existence of two turn-taking systems is the most decisive characteristic of schisming. This view was supported by Egbert, who demonstrated that although schisming is a participation framework with two simultaneous conversations, each with its own turn-taking system, there is an interface between them during schisming [15]. She also makes a systematic differentiation between overlap and simultaneous talk during schisming. In overlap, simultaneous speakers compete for the floor, an event usually resolved by returning to *one-speaker-at-a-time*. In schisming by contrast, simultaneous speakers orient to one of two distinct floors, an event which if resolved successfully, results in the establishment of two floors [15, p. 43]. Overlapping speech is therefore expected to occur throughout the lifespan of all conversation floors within an F-formation.

7.4 Methodology

In this section we build upon the previously discussed concepts to propose using simultaneous speakers in an F-formation as an initial conservative indicator of the existence of distinct conversation floors.

A common concern with observing groups of conversing people is the potential violation of privacy. In our experience with collecting group interaction datasets, participants often regard having their microphone data recorded and transcribed as being more invasive than being captured on video. In these situations, the lack of verbal information makes it extremely challenging to infer the topics being discussed. How can we then investigate the existence of distinct conversations? Two observations could prove useful:

Inferring schisms without audio data. The relationship between body movements such as gestures and speech has been long established in literature [32]. Some works have shown promising results in estimating the presence of voice activity from automated gestural analysis or accelerometer data [33–35]. It therefore seems feasible that speaker turns can be automatically estimated without audio data. Combined with the observation that the co-existence of two turn-taking systems is the most decisive characteristic of schisming,

we argue that it is in turn reasonable to explore the inference of schisms without audio data through speaking turns.

Linking schisming to floors and F-formations. While Egbert does explicitly use the term *floor* to describe the conversations resulting from a schism, it is useful to observe how this relates back to Edelsky's view of floors. Edelsky defined floors in terms of the acknowledged *what's-going-on* within a psychological time space. The object of focus here could either be a topic or some other function. To borrow Goffman's terms, a schism effectively changes a gathering into a *multifocused* one, where each object of focus can be viewed to correspond to a floor in Edelsky's definition. However, if the participant's lower bodies remain configured such that their transactional segments overlap to produce a common o-space, they would still remain in the same F-formation even if the conversation has undergone a schism into two or more distinct floors. Figure 7.1 depicts this situation conceptually.

Combining these two broad observations, we argue that it is feasible to explore the existence of distinct conversation floors within an F-formation without audio data, whilst capturing speaker turns from visual observations. We propose to start with the following metric. Given a sliding window w of speaking duration d, we consider a *speaker* to be a participant who speaks for the entire duration d. The number of simultaneous *speakers* thus defined corresponds to the number of distinct conversation floors at that position of w, since they correspond to speaking turns in distinct floors.

Of course, the metric is inextricably tied to the duration d being considered; too short a duration, and the concurrent turns might capture either backchannels or the overlapping speech within the same floor as described in Egbert's work. However, a reasonably long duration would capture the speaking turns of participants holding distinct floors. This leads to the question: what qualifies as a reasonable choice for d to differentiate overlaps within a floor from turns in distinct floors? In our experiments, we set the lower bound of d at one second. Here we provide evidence from literature to justify this choice.

Choice of speaking window duration. In a study of gaps and overlaps in conversations, Heldner and Edlund report that on average 40% of the speaker transitions in their corpora involved overlaps (including any overlap of over 10 ms) [36]. These represent overlaps for competing for the floor. As for the duration of these overlaps, their histogram makes clear that the duration follows a mode of 50 ms in the Spoken Dutch Corpus, with a mean of 610 ms, and median of 470 ms, all under one second. In a follow-up detailed statistical analysis, Levinson and Torreira differentiate between types of overlaps: *between-overlaps*, that refer to overlaps where the floor was transferred without a silent gap between speakers; and *within-overlaps*, where overlapping speech occurred in between a speaking turn and did

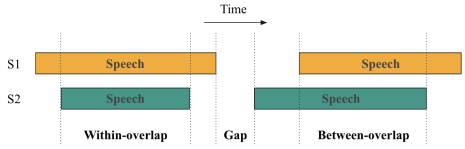


Figure 7.2: Illustration of gaps, within-overlaps, and between-overlaps for two speakers (S1 and S2) within the same floor. The scheme was originally proposed by Heldner and Edlund [36] and adopted by Levinson and Torreira in their analysis [37].

Figure 7.3: Snapshots of the mingling session (Cameras 1-5) in MatchNMingle.

not result in a transfer of floor [37]. Figure 7.2 illustrates these types of overlaps. They used the Switchboard Corpus of English telephone conversations for their analysis, and found that only 3.8% of the signal corresponded to simultaneous speech of both speakers. This fits well with Sacks and colleagues' observations that "overwhelmingly, one party speaks at a time" [13, p. 700], for physically situated embodied social interactions. As for the duration, *between-overlaps* had a modal duration of 96 ms, a median of 205 ms, a mean of 275 ms. On the other hand, *within-overlaps* exhibited an estimated modal duration of 350 ms, a median of 389 ms, a mean of 447 ms. Further, of all the overlaps annotated, 73% involved a backchannel. These statistics indicate that choosing a lower bound for d would reasonably capture simultaneous speech that does not belong to the same floor.

As for the upper bound, a reasonable value should be at least greater than the average turn duration of a speaker. Using the same operationalization proposed in [36], Levinson and Torreira report that contiguous speech delimited by a silent interval of at least 180 ms had a mean duration of 1680 ms, and a median of 1227 ms.

7.5 Dataset

For this study, we use the publicly available *MatchNMingle* dataset [38] that records in-the-wild interactions of 92 people during speed-dates followed by a cocktail party. Three sessions of speed-dates and mingling were recorded in all across three days. We specifically focus on the cocktail party recordings that capture free standing conversations between participants. Figure 7.3 shows the video recordings from five cameras on the last day of data collection. The participants were not given a script to follow and were free to choose the participants they wished to interact with. This allows us to study naturally evolving F-formations and conversation floors in an in-the-wild setting.

Dataset Statistics. The dataset consists of a total of 92 single, heterosexual participants (46 women: 19-27 years with a mean age of 21.6 years and standard deviation of 1.9 years; and 46 men: 18-30 years with a mean age of 22.6 years and standard deviation of 2.6 years).

Over 45 minutes of free mingling interaction were recorded for each of the three days; 56 minutes on the first, 50 minutes on the second, and 45 minutes on the third, respectively.

Annotations. The dataset provides of annotations for both F-formations and a variety of social actions. The F-formations were annotated directly from a video of the interacting participants captured from overhead cameras. The annotations were made for every second for an interval of 10 minutes per day. Each F-formation annotation provides the participant IDs for its members and the start and end times delimiting the lifetime of the F-formation. In all, 174 F-formations were annotated across 30 minutes. Of these, we filtered out those with cardinality less than four, and those for which a participant was found to leave the field of view of the cameras. This left us with 34 F-formations for our experiments.

Of the social actions annotated, we only use the Speaking Status—defined as whether or not a person is speaking. The social actions were annotated for a 30 minute segment for each day, by eight annotators hired for the task and trained by an expert. The annotations were made at the frame level using a tool that allowed for interpolation across frames. In all, 20 annotations per second for each social action are provided. Further, the speaking status is estimated from video alone, by observing lip movements or inferring from the participants' head and body gestures.

7.6 Experiments

We perform two sets of experiments: first we identify the number of simultaneous speakers in an F-formation using the methodology described in Section 7.4, and then evaluate whether the number of members in an F-formation (cardinality) affects the speaking duration of simultaneous speakers.

Simultaneous Speakers in an F-formation. The purpose of this experiment is to evaluate the following—can we infer the existence of distinct conversation floors within an F-formation from simultaneous speaker turns? To recap, this intuition build upon early work on schisming indicating that the co-existence of two turn-taking systems is the most decisive characteristic of distinct conversation floors [13, 14]. Here we consider F-formations of cardinality four and above, since the possibility of distinct conversations occurs only for those F-formations.

We slide a window w of duration d across the lifetime of the F-formation in steps of one second. For every position of w, we count the number of participants with a positive speaking status for the entire duration d. We plot the maximum number of simultaneous speakers over all positions of w. Following the formulation described in Section 7.4, this represents the maximum number of distinct conversation floors that were observed during the life-time of the F-formation. We vary d from 1-20 seconds to guard against the possibility that the smaller values of d might capture co-narration or overlaps within the same floor.

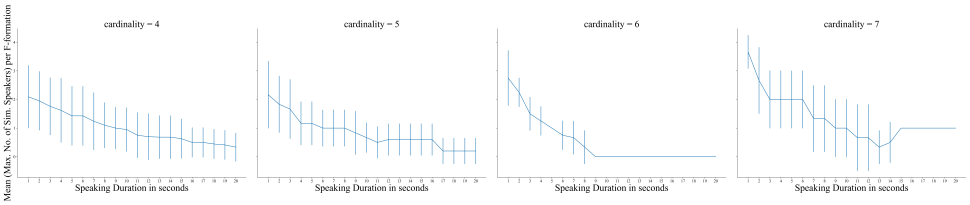


Figure 7.4: Plotting the effect of varying the speaking duration threshold d on the number of simultaneous speakers per cardinality of F-formation. To aggregate the data from each F-formation, the maximum of the number of simultaneous speakers is considered over all the sliding window positions across the lifetime of the F-formation. The y-axis plots the Mean (Maximum number of simultaneous speakers over window positions) over F-formations.

The upper bound of 20 seconds was chosen as sanity check; we expected to see very few speakers have a speaking turn that long.

The *max* operator was chosen to aggregate the number of simultaneous speakers across all window positions into the most conservative measure for what this experiment seeks to evaluate. A value of one for the maximum number speakers over all positions of w would indicate that only a single conversation floor existed within the F-formation. Therefore, observing values greater than one for the *max* metric would indicate the presence of distinct floors with more certainty than other choices of summarizing statistics.

Figure 7.4 plots the mean number of distinct conversation floors per F-formation against varying values of d, per cardinality of F-formation. Cardinality here refers to the number of members in an F-formation. As a sanity check, we would expect the numbers upper-bounded by the number of people in the F-formation; at worst, every person in the F-formation speaks simultaneously to compete for the floor they are a part of. On the same note, we observe that the starting mean values all seem reasonable: about 2 for cardinalities four and five, about 3 for cardinality six, and about 4 for cardinality seven. Assuming that it is common for speakers to have at least one conversing partner, we would expect about half the number of simultaneous speakers as members in an F-formation. Our minimum choice of d was chosen to be greater than the modal duration of overlaps found in previous work [37], so it is less likely that the lower turn durations capture competing overlaps for the same floor. Moreover, at the average turn length of about two seconds observed by Levinson and Torreira [37], we observe that the maximum number of simultaneous speakers is greater than one at all cardinalities considered. This suggests that the simplifying assumption from previous research of a single conversation within an F-formation is insufficient.

We also observe a decreasing trend for the curves in Figure 7.4. This seems intuitive, as it is much less likely that participants would speak for the entire duration of a window as d increases. Interestingly, there is a single example of a speaker speaking for 20 seconds in

an F-formation of cardinality seven. On closer inspection, this turned out to be an error in speaking status annotation, and we manually fixed this error for subsequent analysis.

Effect of cardinality on turn duration of simultaneous speakers. Sacks et al. observed that there is a "pressure for minimization of turn size, distinctively operative with three or more parties" [13, p. 713]. They note that the possibility of a schism introduced by the fourth participant may influence the turn-taking system by 'spreading the turns around' if there is an interest in retaining participants in the conversation. However, they concede that this effect is equivocal, since turn distribution can also be used for encouraging schisming. In this experiment, we explore this effect and pose the question as follows: for a given speaking turn duration d, do we observe a decrease in the maximum number of conversation floors observed over an F-formation's lifetime with an increase in the cardinality of an F-formation?

Qualitatively, this corresponds to the steepness of fall-off of the curves in Figure 7.4. It seems that the the curves for cardinality six and seven falloff more steeply than those for cardinalities four and five. To quantitatively test if cardinality has an effect, we fit a Generalized Linear Model (GLM) to the same data as in the previous experiment with an interaction factor between cardinality and the speaking turn duration d. Specifically, we assume the maximum number of simultaneous speakers observed over the lifetime of each F-formation, y_i to be realizations of independent Poisson random variables, with $Y_i \sim P(\mu_i)$ and model μ_i as follows:

$$\log(\mu_i) = \beta_0 + \beta_1 * d_i + \beta_2 * c_i + \beta_3 * d_i * c_i \tag{7.1}$$

where d_i refers to the duration of the speaking window, and c_i refers to the cardinality for the ith observation. The βs refer to the regression coefficients. The GLM was fit using the *statsmodels* python package. The results of the GLM regression test are provided in Table 7.1. We conclude that cardinality and the two-way interaction between cardinality and turn duration are statistically significant at a significance level of 0.01. Turn duration is itself significant at a significance level of 0.05.

While the previous test tells us that turn duration and cardinality are significant, we still need to perform post-hoc comparisons to ascertain the differences between the cardinalities. We fit multiple GLMs to each possible pair of cardinalities being considered and correct

Table 7.1: Generalized Linear Model Regression Results

	Coef (β)	Std Err	z	P>|z|
Intercept	0.0626	0.339	0.184	0.854
Turn-duration	0.0057	0.002	2.296	0.022
Cardinality	0.1869	0.072	2.603	0.009
Turn-duration:Cardinality	-0.0025	0.001	-4.543	0.000006

Table 7.2: Nominal *P*-values for Six Post-Hoc GLM Regression Comparisons

Cardinality Pairs	Intercept (β_0)	d (β_1)	c (β_2)	d:c (β_3)
4-5	0.196	0.855	0.794	0.403
4-6	0.364	0.0007	0.010	0.00002*
4-7	0.697	0.428	0.030	0.009
5-6	0.079	0.0008	0.016	0.00016*
5-7	0.434	0.413	0.043	0.052
6-7	0.275	0.006	0.657	0.024

d = turn-duration, c = cardinality, d:c = interaction-factor. βs denote the corresponding regression coefficients. * denotes significance at a threshold of 0.001 after Bonferroni correction for six tests.

the corresponding p-values using the Bonferroni correction for multiple testing. Table 7.2 provides the corrected p-values for the post-hoc comparisons. From the last column, we find that cardinality and its interaction with turn-duration are significant between the cardinalities {4, 6}, and {5, 6} at a significance level of 0.001.

One potential limitation of this analysis is the imbalance in the number of F-formations of different cardinalities. F-formations of cardinality four were the most common in the data, with reasonable number of samples to infer a pattern. We believe that the intuition of cardinality and its interaction with speaking turn duration being significant is still a sound intuition, although the statistical significance should perhaps be viewed within the context of the number of F-formations we see in the data. Figure 7.5 plots the number of observations that contributed to the graphs in Figure 7.4.

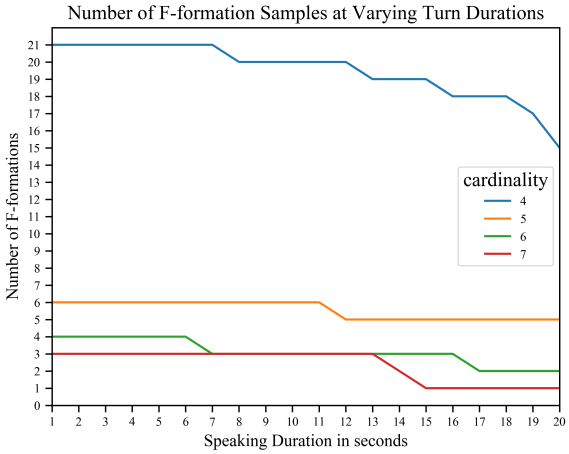


Figure 7.5: Number of F-formations at different speaking turn durations.

7.7 Conclusion

In this study, we presented an initial exploration into unifying the spatial and temporal perspectives of a free-standing conversing group. Specifically, we proposed using simultaneous speaking turns as an indicator for the existence of distinct conversation floors. In the absence of audio data to identify the topics being discussed, our proposed metric can be used to gain a deeper understanding of the conversation dynamics within an F-formation, since speaking turns can be inferred from visual or wearable-sensor data. Our experiments demonstrate that at an average turn duration of two seconds for humans [37], there is evidence of multiple conversation floors within a single F-formation. Further, we found that an increase in cardinality of an F-formation correlates with a decrease in turn duration of simultaneous speakers, specifically between F-formations of sizes {4,6}, and {5,6} in our data. A deeper analysis would be required to identify whether the differences in F-formations of cardinality six hold across datasets, with preferably more examples of F-formations of size six and greater. In this initial approach to the problem, our study does not account for the behaviour of the silent participants, or the evolution of turn taking dynamics within a floor. These remain promising avenues to explore for future works.

Acknowledgments

Chirag Raman thanks Stavros Makrodimitris, Madhumita Sushil, Giovanni Cassani, Erik B. van den Akker, and Yeshwanth Napolean for their time and thoughtfulness.

References

[1] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 2009.

[2] H. Huettenrauch, K. S. Eklundh, A. Green, and E. A. Topp. Investigating spatial relationships in human-robot interaction. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006.

[3] M. Vazquez, A. Steinfeld, and S. E. Hudson. Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation. IEEE, 2015.

[4] M. Vazquez, A. Steinfeld, and S. E. Hudson. Maintaining awareness of the focus of attention of a conversation: A robot-centric reinforcement learning approach. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016.

[5] M. Vazquez, E. J. Carter, B. McDorman, et al. Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*. ACM Press, 2017.

[6] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity

recognition. In *European Conference on Computer Vision*. Springer Berlin Heidelberg, 2012.

[7] K. Tran, A. Gala, I. Kakadiaris, and S. Shah. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. *Pattern Recognition Letters*, 2014.

[8] M. Cristani, L. Bazzani, G. Paggetti, et al. Social interaction discovery by statistical analysis of f-formations. British Machine Vision Association, 2011.

[9] M. Farenzena, A. Tavano, L. Bazzani, et al. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 2013.

[10] L. Bazzani, M. Zanotto, M. Cristani, and V. Murino. Joint individual-group modeling for tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[11] G. Groh, A. Lehmann, J. Reimers, M. R. Friess, and L. Schwarz. Detecting social situations from interaction geometry. In *2010 IEEE Second International Conference on Social Computing*, 2010.

[12] H. Hung and B. Kröse. Detecting f-formations as dominant sets. ACM Press, Proceedings of the 13th international conference on multimodal interfaces, 2011.

[13] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Studies in the organization of conversational interaction*, 1974.

[14] C. Goodwin. Forgetfulness as an interactive resource. *Social Psychology Quarterly*, 1987.

[15] M. M. Egbert. Schisming: The collaborative transformation from a single conversation to multiple conversations. *Research on Language & Social Interaction*, 1997.

[16] A. Kendon. *Conducting interaction: patterns of behavior in focused encounters*. Cambridge University Press, 1990.

[17] C. Edelsky. Who's got the floor? *Language in Society*, 1981.

[18] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani. Multi-scale f-formation discovery for group detection. pages 3547–3551. IEEE International Conference on Image Processing, 2013.

[19] F. Setti, C. Russell, C. Bassetti, and M. Cristani. F-formation detection: Individuating free-standing conversational groups in images. *PLOS ONE*, 2015.

[20] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe. Analyzing free-standing conversational groups: A multimodal approach. ACM Press, Proceedings of the 23rd ACM international conference on Multimedia, 2015.

[21] S. Vascon, E. Z. Mequanint, M. Cristani, et al. Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Computer Vision and Image Understanding*, 2016.

[22] E. Ricci, J. Varadarajan, R. Subramanian, et al. Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015.

[23] H. Hung, D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez. Investigating automatic dominance estimation in groups from visual attention and speaking activity. ACM Press,

International conference on Multimodal interfaces, 2008.

[24] L. Zhang and H. Hung. Beyond f-formations: Determining social involvement in free standing conversing groups from static images. IEEE Proceedings of the Conference on Computer Vision and Pattern Recognition, 2016.

[25] L. Zhang and H. Hung. On social involvement in mingling scenarios: Detecting associates of f-formations in still images. *IEEE Transactions on Affective Computing*, 2018.

[26] C. Oertel, K. A. Funes Mora, J. Gustafson, and J.-M. Odobez. Deciphering the silent participant: On the use of audio-visual cues for the classification of listener categories in group discussions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*. ACM Press, 2015.

[27] C. Oertel and G. Salvi. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *Proceedings of the 15th ACM on International conference on multimodal interaction - ICMI '13*. ACM Press, 2013.

[28] D. Bohus and E. Horvitz. Managing human-robot engagement with forecasts and... *um* ... hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14*. ACM Press, 2014.

[29] A. Pentland, A. Madan, and J. Gips. Perception of social interest. In *Proceedings of the 5th international conference on devlopment and learning ICDL 2006*. Department of Psychological and Brain Sciences, Indiana University„ 2006.

[30] H. Hung and D. Gatica-Perez. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia*, 2010.

[31] E. Goffman. *Behavior in public places: notes on the social organization of gatherings*. The Free Press, 1. paperback ed., 24. printing edition, 1966.

[32] D. McNeill. *Language and Gesture*. Cambridge University Press, 2000.

[33] H. Hung and S. O. Ba. Speech/non-speech detection in meetings from automatically extracted low resolution visual features. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010.

[34] E. Gedik and H. Hung. Speaking status detection from body movements using transductive parameter transfer. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 2016.

[35] E. Gedik and H. Hung. Personalised models for speech detection from body movements using transductive parameter transfer. *Personal and Ubiquitous Computing*, 2017.

[36] M. Heldner and J. Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 2010.

[37] S. C. Levinson and F. Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 2015.

[38] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. v. d. Meij, and H. Hung. The MatchNMingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed years. *IEEE Transactions on Affective Computing*, 2018.

8

Perceived Conversation Quality in Spontaneous Interactions

N. Raj Prabhu, **C. Raman**, and H. Hung. Defining and Quantifying Conversation Quality in Spontaneous Interactions. *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion)*, Virtual Event, Netherlands, 2020, pp. 196-205. DOI: 10.1145/3395035.3425966.

C. Raman*, N. R. Prabhu*, and H. Hung. Perceived Conversation Quality in Spontaneous Interactions. *IEEE Transactions on Affective Computing*, pp. 1-13, 2023. DOI: 10.1109/TAFFC.2023.3233950.

*Equal contribution

Abstract

The quality of daily spontaneous conversations is of importance towards both our well-being as well as the development of interactive social agents. Prior research directly studying the quality of social conversations has operationalized it in narrow terms, associating greater quality to less small talk. Other works taking a broader perspective of interaction experience have indirectly studied quality through one of the several overlapping constructs such as rapport or engagement, in isolation. In this work we bridge this gap by proposing a holistic conceptualization of conversation quality, building upon the collaborative attributes of cooperative conversation floors. Taking a multilevel perspective of conversation, we develop and validate two instruments for perceived conversation quality (PCQ) at the individual and group levels. Specifically, we motivate capturing external raters' gestalt impressions of participant experiences from thin slices of behavior, and collect annotations of PCQ on the publicly available MatchNMingle dataset of in-the-wild mingling conversations. Finally, we present an analysis of behavioral features that are predictive of PCQ. We find that for the conversations in MatchNMingle, raters tend to associate smaller group sizes, equitable speaking turns with fewer interruptions, and time taken for synchronous bodily coordination with higher PCQ.

Index Terms*: Perceived Conversation Quality, Spontaneous Interactions, Social and Behavioral Sciences, Group Interactions*

8.1 Introduction

Picture a spontaneous interaction such as a daily social conversation at work or home. The quality of such conversations is of importance towards both our well-being as well as the development of interactive technologies that influence our daily lives. At an individual level, conversation quality is directly associated with our happiness and life satisfaction [1, 2]. Furthermore, human judgement of conversation quality is a common measure for the evaluation of artificial conversation agents [3, 4]. Despite its importance, little prior research has directly studied conversation quality or jointly considered the factors affecting its perception.

One challenge is that conversation quality is not directly measured, and needs to be inferred from observable verbal and non-verbal behavioral cues. This has led to some research viewing conversation quality in narrow terms, considering only isolated attributes of the conversation. For instance, Milek et al. [1] and Mehl et al. [2] consider greater conversation quality to correspond to less small talk and information exchange at more than a trivial level of depth. On the other hand, taking a broader view of conversation quality runs into another challenge: its potential intersection with several overlapping social concepts. These include rapport [5], bonding [6], interest-levels [7], and involvement [8] amongst others. When studied towards the development of interactive dialogue agents, the focus

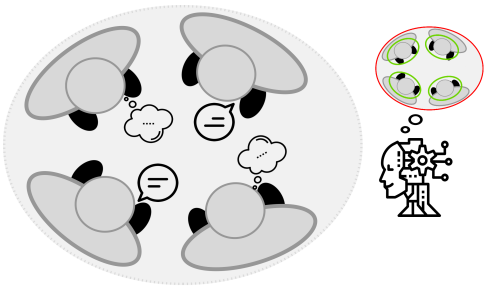


Figure 8.1: Conceptual illustration of individual experiences existing in the perception of interacting partners, and how an external perceived measure of individual-level (green) and group-level (red) experience is relevant for the development of artificial interactive social agents.

has been on the verbal content of non-spontaneous dyadic conversations with a chatbot [3, 4]. In the second ConvAI2 Challenge, the human judgment of quality was evaluated simply as a measure of enjoyment through the question *"How much did you enjoy talking to this user?"* [4]. See et al. [3] conducted a large-scale study to identify the fine-grained factors governing human judgments of full conversations. Even here, the human judgment of overall quality is expressed in terms of the *humanness* and *engagingness* of artificially generated verbal dialogues. Moreover, the recording of spontaneous conversations in a way that enables the transcription of verbal content constitutes a privacy concern with ethical implications [9, 10]. Consequently, while individual factors have been studied in isolation, joint consideration of the multiple aspects of conversation quality in natural, spontaneous conversations remains a knowledge gap.

In this work, we take the perspective that such a *holistic* characterization of the quality of multiparty spontaneous interactions is an important objective in the development of socially intelligent systems. For instance, consider a social robot approaching a conversing group of people, as illustrated in Figure 8.1. Here, a perception of the group's experience of the conversation as a whole could aid the social agent in developing more nuanced policies of approach. Furthermore, an estimate of each individual's experience could then aid the agent in developing personalized adaptive strategies to conduct the subsequent interaction smoothly.

In addition to a holistic characterization, we specifically argue for a *perceived* measure of conversation quality in this work, at both the individual and the group levels. This is in contrast to existing efforts for quantifying quality-related aspects of conversations, which have largely focused on self-reported measures after interactions [5, 6, 11, 12]. While such measures attempt to estimate an individual's true experience in situ, they also suffer several drawbacks including desirability bias [13], egoistic bias [13, 14], and recall bias and cognitive errors [15]. On the other hand, a perceived measure of experience quantifies how

participants seem to be experiencing the interaction to an external third-party observer [7, 8, 16]. While such a measure may not capture the true experience, it closely models how we conduct interactions based on imperfect estimates of our conversation partners' experiences, and is therefore also useful towards the development of machines with social intelligence.

Concretely, we make three contributions in this work. First, we introduce the novel measure of *Perceived Conversation Quality* (PCQ) towards quantifying social experience in spontaneous interactions by jointly considering potentially overlapping related constructs. Second, we present an instrument for collecting annotations of PCQ at both the individual and the group level. We validate the instrument on the publicly available MatchNMingle dataset [9] of mingling interactions following a speed-dating event. Third, we present insights into the behavioral features that predict PCQ through confirmatory statistical analysis and empirical data-driven analysis.

Our preliminary work on this topic was presented in [17], which described the proposed instrument and analysis of annotations. The experiments we present in this manuscript (Section 8.5 onward) are completely new. Moreover, this manuscript is a complete rewrite; compared to our prior publication the manuscript now includes a clearer (i) overall presentation and motivation, (ii) organization of related literature, and (iii) description of the process of conceptualizing, validating, and analyzing PCQ.

8.2 Related work

Spontaneous interactions are considered to be non task-directed, unconstrained, and typically occurring in natural situations [18–20]. In such a dynamic conversation setting, several constructs emerge. These include descriptors of interpersonal relationships amongst participants (e.g. rapport [5] and bonding [6]), or those which capture qualitative attributes of the interaction (e.g. involvement [8, 21], engagement [22], and interest-levels [7]).

8.2.1 Rapport and Bonding

Rapport and bonding have been widely studied as a pairwise phenomena using self-reported measures [5, 6, 11]. Müller et al. [5] define rapport as "the close and harmonious relationship in which interaction partners are 'in sync' with each other". The authors used a self-reported questionnaire adapted from Bernieri et al. [23] to measure rapport for every pair of individuals within small interaction groups. Another related social concept is bonding, which measures positive personal attachment including "mutual trust, acceptance, and confidence" amongst interacting pairs [24]. Based on this definition, Jaques et al. [6] studied bonding in human-agent interactions, using the bonding subscale of the *Working Alliance Inventory* (B-WAI) [24].

8.2.2 Involvement, Engagement, and Interest-Levels

Antil [21] defines involvement as "the level of perceived personal importance and/or interest evoked by a stimulus (or stimuli) within a specific situation". Following Antil's view of involvement as a non-binary variable, Oertel et al. [8] developed a 10-level annotation scheme for joint involvement of a group based on intuitive, listener-independent impressions of prosody and body and face movement. Oertel and Salvi [25] proposed a gaze-based method to relate group involvement to individual engagement in multiparty dialogue. Several researchers have conceptualized group cohesion to study its influence on task performance [26], in settings such as meetings [27, 28] and long-term crew missions [29, 30]. Gatica-Perez et al. [7] define group interest-levels as, "the perceived degree of interest or involvement of the majority of the group". The authors provided perceived annotations for interest-levels using audio-visual recordings of interactions, on a discrete 5-point scale. To this end, the external annotators were instructed to attend to interest-indicating activities such as note-taking, focused gaze, and avid participation in discussion. Note that these constructs have all been defined and studied in task-directed settings.

8.2.3 General Measures of Interaction Experience

In contrast to efforts focusing on specific social concepts, some recent approaches have proposed more general measures of experience in conversations. Cuperman and Ickes [12] introduced the *Perception of Interaction* (POI) questionnaire as part of a study to examine the effects of gender and personality traits on participant behaviors in dyadic interactions. The questionnaire collected self-reported measures of a participant's perception of their interaction experience. These aspects included the perceived quality of the interaction, the degree of rapport they felt they had with the other person, and the degree to which they liked the other person. This measure of interactions has been adapted by other works to study bonding [6] and interaction experience [31]. Lindley and Monk [16] follow the rationale that experience itself is difficult to quantify, but since it is entwined with social interaction, we might characterize experience by measuring aspects of conversation that are related to it. They studied several behavioral process measures and developed the *Thin-Slice Enjoyment Scale* (TES): a measure of empathized enjoyment in social conversations from ratings of thin slices of behavior by naïve judges. In their factor analysis, the authors found that the judges viewed enjoyment and conversation fluency as being related. However, the POI was developed for self-reported measures, and neither work considered spontaneous interaction settings: Cuperman and Ickes [12] considered scripted dyadic interactions with confederates, while Lindley and Monk [16] developed the TES within the particular task-directed context of photo sharing.

8.3 Perceived Conversation Quality

8.3.1 Initial Conceptualization

The primary influences for our conceptualization of PCQ are the works of Edelsky [32], Lindley and Monk [16], and Cuperman and Ickes [12]. Specifically, from these works we motivate the rationale behind our choices of (i) focusing on the cooperative aspects of conversation towards conceptualizing PCQ, and (ii) rating thin slices of behavior to capture the gestalt impressions raters have of the continually unfolding conversation.

In an analysis of social interactions in a series of meetings, Edelsky [32] observed two contrasting styles of conversation, termed *cooperative floors* and *exclusive floors*. Cooperative floors are characterized by collaborative stretches of "free-for-all" conversation accompanied by a feeling of participants being "on the same wavelength" [32, p. 391]. (In contrast, the exclusive floor is owned by a single person with turns rarely overlapping.) This notion of the cooperative floor captures the sense of engagement associated with positive experiences, and has been since linked with informal social interactions [33–35] and enjoyment [36]. As such, we observe that Edelsky's notion of "on the same wavelength" strongly resonates with the POI questionnaire's focus on how interaction partners relate to each other [12]. Subsequent researchers have also derived qualitative measures of conversation based on the "free-for-all" aspects of Edelsky's description. These include conversational equality and freedom [16] (or interactivity [37]), and fluency through the occurrence of frequent turns [16, 38].

Ambady and Rosenthal [39] propose that thin slice judgments of behavior can be usefully made so long as the variables in question are observable and there is an affective or interpersonal component. They suggest that this is because such inferences are made through subconscious decoding of expressive behavior, with judgemental accuracy being strongly linked to "gestalt, molar impressions based on nonverbal behavior" [40, p. 439]. This result supports previous research showing that molar impressions, although vaguer and fuzzier, generally yield more useful information than the coding of specific behaviors without accounting for overall context. Researchers often encourage the formation of this gestalt impression by intentionally reducing information presented to raters, e.g. removing speech content while retaining tone of voice or extinguishing facial expressions [41]. In contrast, obtaining judgments of gestalt impressions is a natural fit for spontaneous interaction settings where recording speech or ego-centric perspectives is often not possible to preserve privacy [9, 10, 42].

8.3.2 Pilot Qualitative Interviews with Naïve Judges

We conducted pilot qualitative interviews with three naïve judges to verify if our initial conceptualization matched the lay interpretation of PCQ. All judges were students enrolled

Figure 8.2: A snapshot from the MatchNMingle dataset [43].

in technical Masters programs at the authors' university. The judges were shown unaltered recordings from the publicly available MatchNMingle (MnM) dataset [43], and asked what they thought of the conversations in the scene. Figure 8.2 illustrates a snapshot of a scene from MnM. To obtain unbiased impressions, we didn't specify our focus on conversation quality, nor our conceptualization of it. All judges (i) described a continually evolving perception of participant experiences over the conversation lifetime, aligning with our choice of rating thin slices of behavior rather than a single rating for the entire conversation; (ii) described perception of individual experiences as well as the group as a whole, aligning with our choice of measuring PCQ at the individual- and group- levels separately; and (iii) identified the attributes of equal opportunity for speaking, smoothness of interaction, and interpersonal relationships that strongly resonates with the prior work that serves as our primary influences [12, 16, 32].

8.3.3 Definition and Constituents

Following our initial conceptualization and pilot interviews, we formalize PCQ of a spontaneous interaction as

> *the degree to which participants in the spontaneous interaction appear to be on the same wavelength and maintain an equal opportunity floor, as perceived by an external observer.*

Further, in the following subsections we present three constituents of PCQ that categorize the multiple social concepts associated with this definition.

Interpersonal Relationships

This constituent describes the degree of association between participants or the notion of being in-sync with one's interaction partners, using constructs such as rapport [5] and bonding [6]. More specifically, the constituent measures the degree to which an individual was accepted and respected by other individuals in the group or the degree to which the other individuals were paying attention to the individual. Increased bonding and rapport

amongst interacting partners is widely acknowledged to result in improved collaboration, and improved interpersonal outcomes, thereby having a key influence on the PCQ.

Nature of Interaction

This constituent describes the degree to which the interaction was smooth and relaxed or forced and awkward. It captures the notion of whether the participants are having a positive and pleasant experience, drawing upon the quality of interaction aspects of the POI [12].

Equal Opportunity

This constituent captures the *free-for-all* collaborative aspects of Edelsky's description of cooperative floors [32]. It describes the notion of equality of opportunity for participation shared amongst interacting partners, capturing the sense of cohesiveness and engagement in informal conversations. This includes factors such as conversation freedom [44], equality, and fluency [16] and an individual's opportunity to take the lead in the conversation [6, 12].

8.3.4 PCQ Questionnaires: A Multilevel Perspective

We devise two independent questionnaires to measure PCQ at the individual and group levels. This follows our broader multilevel perspective [45] of social interactions where constructs can be conceptualized at different levels, such as the individual, dyadic, and group levels. While prior works have often considered constructs at a single level (e.g. Müller et al. [5] consider rapport as a dyadic pairwise construct), a multilevel perspective aligns better with our pilot judges' descriptions of attributes pertaining to individuals and groups as a whole. Moreover, some prior works on conversation group dynamics have indeed also taken a multilevel perspective: Oertel and Salvi [25] distinguish overall group involvement from individual engagement, obtaining separate annotations at both levels. In the case of PCQ, our view is that an observer's perceptions of individual affect and behavior dynamically interact to contribute to an overall group-level perception. Figure 8.3 illustrates the scope of observations towards measuring PCQ at each level.

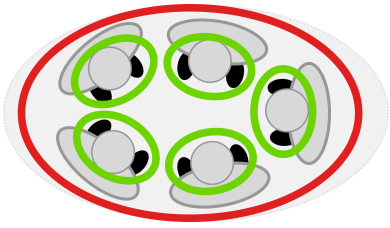


Figure 8.3: Illustrating the scope of observation to measure the group-level (red) and individual-level (green) PCQ.

The individual level captures what the quality of the conversation appears to be to a particular individual. The focus is on how the individual seems to be relating to their partners and participating in the conversation. Consequently, every individual receives a rating. Note that this perspective doesn't consider the individual's behavior in *isolation* by excluding the context of partner behaviors. Rather, the scope of consideration is restricted to what the individual seems to be experiencing. In contrast, the group level expands this scope of consideration to all interlocutors *as a whole*, focusing on their collective experience, resulting in a single group-level rating.

Concretely, we devise the PCQ questionnaires by drawing upon elements of the POI scale [12] and the TES [16]. However, since the POI was developed for self-reports rather than external perception, and neither was developed for spontaneous interaction settings, we adapt the specific items. First, all items were updated to address external observers and apply to group sizes beyond dyads. Second, privacy-preserving datasets of in-the-wild conversations often omit recording audio. So items referring to the verbal or paralinguistic content of speech were skipped, thereby relying solely on nonverbal cues for perception. Finally, we excluded original items that would require external raters to make significant speculations about participants' desires and opinions beyond what can be inferred from their observable behavior. These include questions related to interpersonal liking (e.g. *"I would like to interact more with the partner in the future"*), or degree of rapport (e.g. *"I felt that the partner was paying attention to my mood"*). From the varied descriptions of pilot judges on the matter, as well as internal author discussions, we deemed that answering such questions require external observers to make too many unverifiable assumptions for a useful perceived measure of conversation quality. We provide the two PCQ questionnaires in Supplementary Material Section 8.A.

8.4 Annotations, Validity, and Reliability

8.4.1 Dataset

We use the publicly available MnM dataset [43]. MnM is a multimodal dataset of in-the-wild free-standing mingling interactions. The recordings constitute a total of 30 minutes of interaction across three days, annotated for conversation groups using the spatial positions of the participants in video from overhead cameras. Figure 8.2 illustrates a snapshot from the dataset. Conversation groups were operationalized using the framework of F-formations [46], where a unique group was considered to be an F-formation with a fixed number of interlocutors. The leaving or joining of one or more members was considered to give rise to new unique conversing groups. The authors of the dataset chose specific windows of 10 minutes per day for annotation with an aim to eliminate possible effects of participant acclimatization to being in a recorded mingling setting, and to maximize the

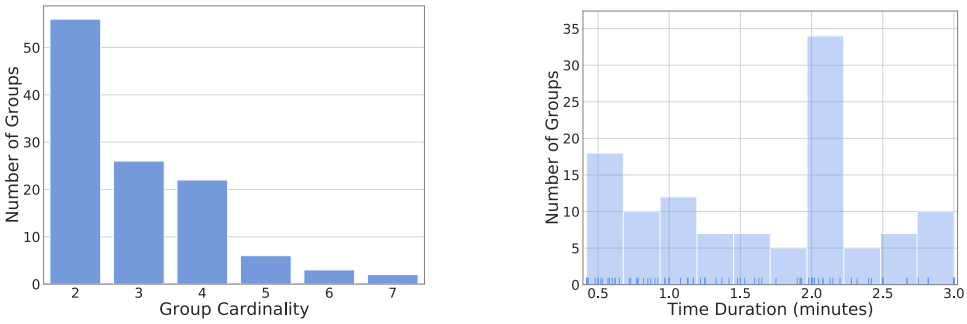
density of participants in the scene. Over the 30 minutes 174 conversation groups were annotated. The duration of group conversation follows a mean of 1.91 min, std. of 2.13 min, median of 1.10 min, and a mode of 0.52 min. The provided data contains video from three of the five overhead cameras, and accelerometer readings from a sensor pack worn by each participant.

8.4.2 Annotation Procedure

The PCQ annotations were performed by only relying on overhead cameras *videos*. The MnM dataset contains only general audio from the overhead cameras, which is insufficient to reliably infer verbal cues of an individual, and close-talk microphone recordings are not available. However, the MnM dataset contains video recordings that capture rich non-verbal behaviors of participants from which a useful perception of conversation quality can be formed [7, 16].

We began by splitting the group conversations into multiple thin-slices [6, 47]. The distribution of group interaction duration in the data follows a median of 1.10 min and a mean of 1.91 min. For a fair comparison to conversations lasting around 1 minute, we split conversations of duration greater than 2 minutes into independent slices of 1 minute each. Conversations of duration less than 2 minutes were untouched. We also omitted groups with a duration of less than 30 seconds. Note that studies on the predictive validity of thin slices of nonverbal behavior for other tasks have revealed (i) no clear pattern for optimal slice locations for 1 min slices within a longer slice [48]; and (ii) only some loss in predictive capacity for 1 min slices, while slices of duration 2 or 3 min were in general equal to 5 min slices in predictive capability [48, 49]. Considering these results along with the distribution of conversation duration in our data, we believe our choice of splitting conversations larger than 2 minutes into 1 minute slices to be reasonable. After the omission of groups lasting under 30 seconds, the total number of resulting conversation groups was 115. The distribution of group cardinality (number of participants) and interaction duration can be seen in Figure 8.4a and Figure 8.4b respectively.

We began by first conducting a qualitative annotation pilot with the same naïve judges who participated in the qualitative interviews. Note that these judges were not used for the final annotations. The goal of this pilot was to fine-tune the final annotation process using any initial feedback about the annotation procedure. The pilot annotators were presented with the videos of the individual thin-slices and asked to fill the two PCQ questionnaires. However, post-hoc interviews revealed two considerations. First, the annotators found the presence of free-standing conversation groups (FCGs) other than the one under consideration distracting. Second, the annotators suffered from fatigue while annotating longer conversations, especially while annotating both individual and group level PCQ. In light of this feedback, we manually cropped each FCG from the overhead



(a) Group Cardinality

(b) Duration of interactions

Figure 8.4: Distribution of conversation group attributes from the MatchNMingle dataset.

video. To further reduce fatigue, annotators were given a period of two months to annotate all the slices, and were instructed to not annotate more than three groups per day.

The final annotations[1] were performed on a 5-point scale by three annotators. The annotators were chosen to be naïve judges in order to capture a general perception of conversation quality. The annotators were aged between 22 and 30 years, 2 females and 1 male. The age range matches overlaps with the reported age range of the participants in the data (18 – 30) [9]. One of the annotators spent time internationally as a Masters student, matching the demographics of the participants. All annotators had completed education at least the Bachelors level. The annotators were provided with the independent conversation slices of cropped video clips and asked to fill out both PCQ questionnaires. The slices were provided to the annotators in randomized order for each annotator, to prevent any annotator bias which might occur from a chronological ordering of the clips.

8.4.3 Validity

When measuring intangible constructs such as PCQ, it is important to assess the validity [50, 51] of the proposed instrument. Broadly, validity deals with whether the instrument indeed measures what it claims to be measuring.

Face Validity

First we tested the face validity of our questionnaire items. Face validity is a consensus measure, and is checked to ensure that the raters accept the instrument [50]. This is done by asking the raters if the items seem valid. Both questionnaires passed the face validity test with full consensus.

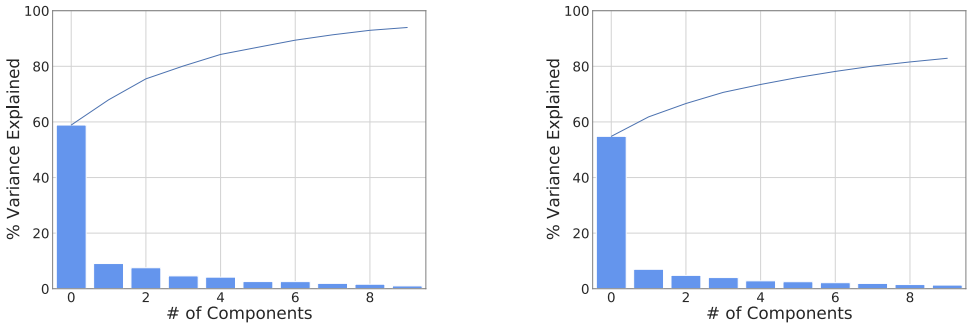
[1]Annotations will be available on the MatchNMingle website at http://matchmakers.ewi.tudelft.nl/matchnmingle/pmwiki/

Criterion and Construct Validity

When prior trusted standards exist for a construct, a criterion-oriented study is common. Here validity can be established by showing that results of administering the instrument correlates with a contemporary criterion (e.g. a psychiatric diagnosis) or by proposing one instrument as a substitute for another (e.g. a multiple-choice form of spelling test is substituted for taking dictation) [51]. However, since PCQ is a novel conceptualization, prior trusted standards do not exist for it. In such cases where the attribute being measured is not "operationally defined", construct validity must be investigated [50, 51]. Construct validation is the gathering of evidence to support the interpretation of what a measure reflects, and addresses the question "What constructs account for variance in test performance?"

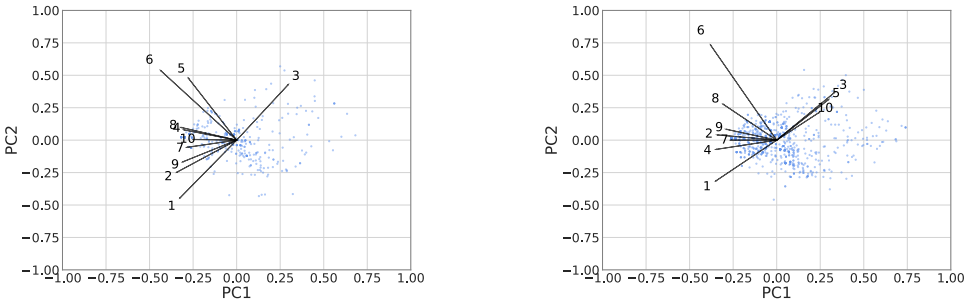
A typical approach for construct validation involves performing a factor analysis and investigating if items corresponding to one construct correlate with each other along a factor (convergent validity) and divert from items of other constructs (divergent validity) [50]. This works well for instruments with independent constructs (e.g. *gender* and *complexity of use* in Brinkman's mobile phone design questionnaire [50, Table 9]). However, such an analysis is unsuitable for situations like ours with overlapping constructs. Indeed, Cuperman and Ickes [12] decided to not reduce items from the POI to a smaller set of factors, following a precedent set by [52]. In contrast, we do perform a factor analysis, but rather than seeking the independence of factors, we investigate whether the loadings correspond to interpretable attributes of the constructs.

A principal component analysis (PCA) of the annotations showed that 71% and 65.2% of the variance at the group-level and individual-level respectively could be explained by the first principal component (see Figure 8.5). Here, 1020 (3*340) and 345 (3*115) *thin-*



(a) Group-level **(b)** Individual-level

Figure 8.5: Eigenvalue distribution (bar chart) and the cumulative percentage of the explained variability (line plot).



(a) Group-level

(b) Individual-level

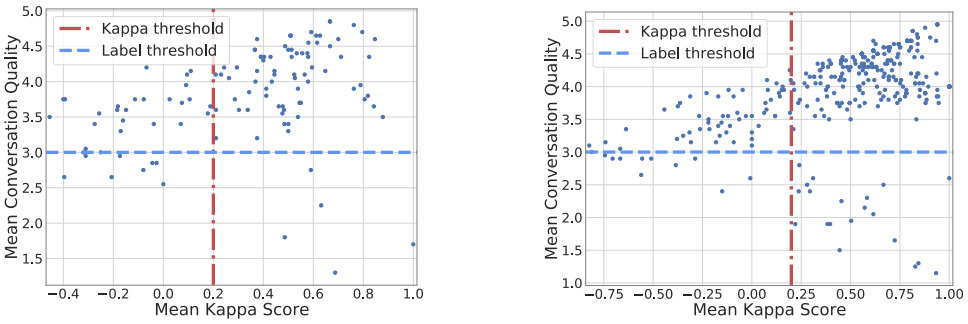
Figure 8.6: Plot of the factor loadings (black lines) and the samples (blue dots) in the first two principal components.

slice samples were used for individual and group level PCQ (i.e., annotations from three annotators for each sample), respectively, with 10 features (the number of questionnaire items), which is greater than the variables-to-features ratio suggested to perform PCA [53]. From the plot of the data samples using the first two principal components in Figure 8.6, we see that questions corresponding to positive and negative orientations of PCQ cluster in opposite directions along the two components. Specifically the individual-level items pertaining to awkwardness (3), discomfort (5), and self-consciousness (10) load in the exactly opposite direction to the item about the individual looking relaxed (1). Of these, at the group-level only items 1 and 3 apply, and we see a similar pattern. Further, we also observe that the items pertaining to *equal opportunity* cluster separately: these correspond to items 5 and 6 about free-for-all participation at the group-level, and item 6 about taking lead at individual-level. Specifically, the highest loading of individual-level item 6 suggests that the taking lead in conversations accounts for the highest variance between individuals, which is intuitive given prior work on dominance in groups [54].

8.4.4 Reliability

To estimate inter-annotator agreement, we use the quadratic weighted kappa measure (κ) [55], a variant of the Cohen's kappa. The measure is especially useful when the annotation data is ordinal in nature. Figure 8.7 plots the mean kappa score against the mean conversation quality score in a scatter plot similar to the analysis of inter-annotator agreement for cohesion performed by Hung and Gatica-Perez [27].

From the plots we see that there exists a linear relationship between mean kappa scores and mean conversation quality scores, suggesting that annotators agree better on conversations of higher perceived quality than conversations of lower perceived quality. Moreover, in the individual-level annotations, there exists a small cluster of samples where



(a) Group-level

(b) Individual-level

Figure 8.7: Scatter plot of the Mean Kappa score (κ) vs the Mean Conversation Quality score.

annotators tended to agree higher for lower conversation quality samples as well. In contrast, annotators never agree well for low conversation quality samples at the group-level.

To handle low inter-annotator agreement, following suggestions by Ringeval et al. [56], we performed zero-mean local normalization to remove annotator bias. Hung and Gatica-Perez [27] omit samples below $\kappa = 0.3$, and Ringeval et al. [56] obtain an average κ of ≈ 0.2 for all their emotion dimensions. Following these approaches, data samples at both the group- and individual- levels with $\kappa < 0.2$ were omitted from further analysis, where a $\kappa > 0.2$ indicates a reliability of *fair and above* [57].

8.5 Modeling Conversation Quality

In this section we describe the experimental setup for our study of behavioral features that can be predictive of PCQ.

8.5.1 Preprocessing

We first preprocess the raw tri-axial acceleration signal from the wearable sensors to extract low-level features. First, each axis recording from the tri-axial accelerometer is standardized by calculating the z-score for each individual and axis, thereby removing the individual differences in movement intensity. Following prior work using wearable sensor data to study conversation dynamics [58–60], we compute the following features using the z-scores: the raw and absolute values for 3 axes each, and the Euclidean norm of the raw values across axes, resulting in a total of 7 feature channels. Further, similar to [60], using a sliding-window filter, we denoise the feature channels by extracting statistical (mean, median and variance) and spectral features (log-bin values of power spectral density) from the respective sliding-windows. Drawing inspiration from [61], we also include features

Table 8.1: An overview of the four sets of individual- and pair- level behavioral features extracted.

Attribute Category		Attribute Variant
Synchrony		
1	Correlation	correlation coefficient (ρ_{xy})
2	Time-lagged Correlation	min, max, argmin, argmax
3	Mutual Information	min, max, mean, variance
4	Mimicry	lag_min, lag_max, lag_mean, lag_variance, lead_min, lead_max, lead_mean, lead_variance
Causality		
5	Coherence	min, max
6	Granger's Causality	f_value
Convergence		
7	Symmetric Convergence	ρ
8	Asymmetric Convergence	lag, lead
9	Global Convergence	$d_1 - d_2$
Turn-Taking		
10	Conversation Equality	degree of equality
11	Conversation Fluency	percentage of silence, # back-channels
12	Conversation Synchronization	percentage of overlap, # successful interrupts, # unsuccessful interrupts

that are not preprocessed to circumvent any data loss from preprocessing. An analysis is also presented to understand their respective benefits (see Section 8.6.2).

8.5.2 Feature Extraction

Individual and pairwise features

We consider *pair-wise* bodily coordination features and *individual-level* turn-taking features to study PCQ. For bodily coordination, we extract three sets of features: *synchrony, convergence,* and *causality.* An overview of the individual and pairwise features extracted can be seen in Table 8.1.

Synchrony. Synchrony estimates the dynamic and reciprocal adaptation of the temporal structure of behaviors between interlocutors [62]. Following existing literature [11, 28, 60], we extract four unique measures of interpersonal synchrony: *Correlation*, *Time Lagged Correlation*, *Mutual Information*, and *Mimicry.* See Supplementary Section 8.B.1 for feature extraction details.

Causality. Correlation does not adequately capture the causal effect [63]. We therefore extract two causality features: *Coherence* [64] and *Granger's Causality.* See Supplementary Section 8.B.2 for feature extraction details.

Convergence. These features capture the increasing similarity between interacting partners over time [65], and have been shown to be predictive of mutual liking, attraction

[60, 66], and social cohesion [28]. In this research, we use three unique estimates of convergence: *Symmetric Convergence*, *Asymmetric Convergence*, and *Global Convergence*. See Supplementary Section 8.B.3 for feature extraction details.

Turn-Taking. MnM provides binary speaking status of participants annotated from video data. We extract turn-taking features using these annotations by assuming a speaking turn to be a continuous speaking activity segment separated by at least 500 ms of silence [16, 67]. Following existing literature [16, 27, 67], we extracted turn-taking features under three categories: *Conversation Equality*, *Conversation Fluency*, and, *Conversation Synchronization*. Assuming a conversation of duration T and a group of N people, and denoting the i-th individual's binary speaking status as $\boldsymbol{s}^i = [s_1^i, \ldots, s_T^i]$, we have the percentage of speaking duration for i, $d^i_{\text{speak}} = (\sum_{t\in[T]} \boldsymbol{s}_t^i)/T$. The degree of equality for i is $eq^i = (d^i_{\text{speak}} - \bar{\boldsymbol{d}})/\bar{\boldsymbol{d}}$, where $\bar{\boldsymbol{d}} = (\sum_{i\in[N]} d^i_{\text{speak}})/N$. As measures of fluency, we compute the percentage of individual silence $d^i_{\text{silence}} = 1 - d^i_{\text{speak}}$ and the number of back-channels (very short utterances of duration up to 2 s). As a measure of synchronization, we consider the percentage of speech overlap, which is $d^i_{\text{o}} = (\sum_{t\in[T]} \mathbb{1}\{s_t^i = s_t^{j,j\neq i}\})/T$ for individual i, and the number of successful and unsuccessful interruptions, which are overlap durations when *turn-change* occurs and does not occur, respectively.

Group-level features

Following [28, 30], we translate individual and pairwise features to group-level features using the feature aggregates *minimum*, *maximum*, *mean*, *mode*, *median* and *variance*. Specifically, for individual-level modeling, similar to Müller et al. [5] we aggregate over pairwise features involving that particular individual, and for group-level modeling aggregation is done over all the pairs in the group.

8.5.3 Experimental Setup

Statistical Analysis

We perform hypothesis-driven tests to study the effect of (i) group cardinality, (ii) turn-taking attributes and (iii) body coordination attributes on PCQ. We use the Quantile Least Squares (QLS) and Joint LASSO models for our hypothesis-driven analysis. The QLS analysis considers each set of behavioral features independently, while the Joint LASSO analysis accounts for the combined effect of all features, allowing for complementary insight. Due to the superior performance of models when no preprocessing was used (empirically explained in Section 8.6.2), for the statistical analysis tests, we only used the features without preprocessing.

Quantile Least Squares. QLS fits the regression to the conditional *median* of the dependent variable, in contrast to the conditional *mean* estimated by Ordinary Least Squares

Table 8.2: Overview of the statistical analysis performed.

Dependent Variables	Independent Variable Sets	Statistical Models
IndivPCQ	Group cardinality	QLS Regression
GroupPCQ	Turn-taking, Bodily Coordination	LASSO Regression

(OLS). Intuitively, the conditional median is more robust against outliers. More crucially, the QLS does not require the data to abide the assumptions of exogeneity and homoscedasticity like the OLS does. We find that the variance of the independent variables varies largely across quantiles (see Supplementary Figure 8.16 for scatter-plots), thereby violating the exogeneity and homoscedasticity assumptions. We therefore use the QLS model for our analysis.

Joint LASSO. While QLS is convenient in situations where classical parametric assumptions do not hold, it still suffers from effects of multicollinearity. We therefore use the QLS model to only study behavioral feature sets in isolation. However, to also account for the combined effect of feature sets, we perform a *joint* regression over all features using a LASSO model, which uses the *coordinate descent* [68] to fit the coefficients, thereby inducing sparsity to address multicollinearity. Subsequently, we perform a post-hoc Spearman's rank correlation on the LASSO filtered features.

An overview of the statistical tests performed can be seen in Table 8.2. We denote individual- and group- level PCQ as IndivPCQ and GroupPCQ respectively. In total, with two dependent variables, three sets of independent variables and three statistical models, 18 tests were performed. Bonferroni correction is applied to the p-values to correct for multiple testing for each dependent variable. After Bonferroni correction a significance threshold of 0.005 was used for testing significance in all the analyses presented.

Analysis of Feature Extraction and Fusion

We perform data-driven analyses to study the effects of (i) window sizes for data preprocessing; (ii) fusion of attribute categories; and (iii) feature aggregators to compute group-level features from individual and pairwise features.

For these analyses, we treat predicting PCQ as a binary classification of low and high PCQ scores. A threshold of 3.0 (on the 5-point scale) is used to binarize the scores into low and high. As such, our annotations suffer from class imbalance, see Figure 8.7 for the label threshold, and Supplementary Section 8.C for the class distribution. To address this, we employ the Synthetic Minority Oversampling technique (SMOTE) [69], which generates synthetic samples from the minority class. We use a logistic regression model trained with the elastic loss that combines the L_1 and L_2 penalties of the lasso and ridge regularization methods. Specifically, for each experiment we evaluate how the feature

extraction or aggregation affects the predictive capability of the model. For dimensionality reduction, we perform PCA on the z-score standardized features, by selecting features that preserve the top 90% of variance in respective predictive tasks. As the performance metric, we use the area under the ROC Curve (AUC) score. The metric is calculated as the average across 5-folds in the cross-validation (CV) setting. A stratified k-fold CV was used to preserve the percentage of samples of each target class as the complete set, in the train and test partitions. Except when studying the effects of preprocessing, in all other experiments only features that are not pre-processed were used. Code for all experiments and analyses are available at https://github.com/LRNavin/conversation_quality.

8.6 Results

8.6.1 Statistical Analysis

Analysis of Group Cardinality

Existing research [70–72] has shown that behavior in group interactions varies with size of the group (group cardinality). Is this true for PCQ as well? We test the hypothesis:

For an FCG, the PCQ changes with group cardinality.

From the plots in Figure 8.8, we see that for both GroupPCQ and IndivPCQ the means for cardinalities of 2, 3 and 4 are higher than that of 5, 6, 7. The statistical tests reveal that IndivPCQ and GroupPCQ are significantly different across groups of different cardinality. We note that for all regression models, the β coefficient for the group cardinality variable is negative, suggesting that PCQ is inversely proportional to group cardinality. For example, the QLS model associates the cardinality attribute with $\beta = -0.2167$ and $\beta = -0.0833$ for IndivPCQ and GroupPCQ respectively (p-value=10^{-5}), indicating that people appear to have better quality conversations with fewer partners.

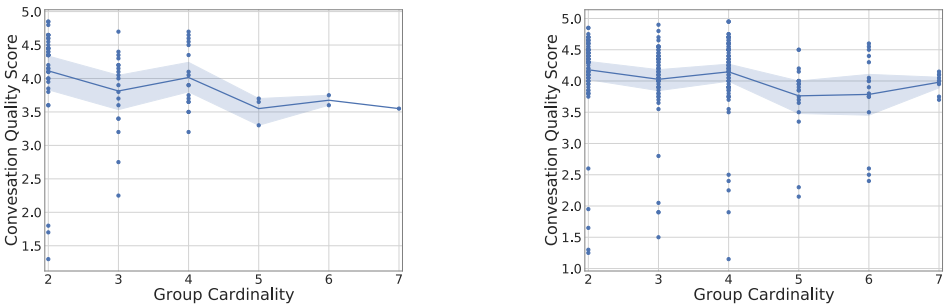


Figure 8.8: GroupPCQ and IndivPCQ across group cardinalities.

Post-hoc analysis testing for the differences in PCQ between cardinality pairs reveals that the IndivPCQ scores are significantly different in dyadic group interactions when compared to that of interactions in larger groups (cardinality ≥ 3). One possible alternate explanation of this result is that raters score PCQ more conservatively when there are more partners to pay attention to. Nevertheless, even if this were the case, it would be a valid characteristic of how people perceive behaviors in larger groups. Significant results were not observed for the post-hoc GroupPCQ comparisons, suggesting that no conclusions can be drawn with respect to GroupPCQ regarding pairwise differences with cardinalities. Note that this result should also be interpreted accounting for the small sample size for cardinalities ≥ 5.

Analysis of Turn-Taking Attributes

Turn-taking features have shown to be indicative of constructs such as enjoyment and cohesion [16, 27, 73]. We test the hypotheses:

> *In an FCG, turn-taking attributes (conversation equality, conversation fluency and conversation synchronization) are positively correlated with PCQ.*

For IndivPCQ, the QLS model reveals that conversation equality and percentage of silence are the most significant attributes, with positive ($\beta = 0.2136, p = 10^{-4}$) and negative ($\beta = -0.5094, p = 10^{-4}$) correlations respectively. For GroupPCQ, QLS reveals that the number of successful and unsuccessful interruptions are the most significant attributes, with negative ($\beta = -0.0859, p = 0.001$) and positive ($\beta = 0.0956, p = 0.002$) correlations respectively. On the other hand, the LASSO and rank correlation models reveal a different set of significant attributes. For IndivPCQ, along with conversation equality and percentage of silence, the two interruption based attributes were also revealed to be significant. Similarly, for GroupPCQ, unlike the QLS, the two interruption attributes are found to be insignificant, while conversation equality, percentage of silence and number of backchannel attributes are found to be significant.

Intuitively, the result implies that observers consider group conversations with more equitable speaking turns and fewer interruptions to be of higher quality. An important thing to note here is that the complementary models associate all attributes with similar trends even though they differ on which attributes they consider to be of statistical significance. Even though the statistical significance of successful and unsuccessful interruptions differ when considered in isolation or jointly with other features, they are associated with negative and positive β's respectively, by all models tested.

Analysis of Bodily Coordination Attributes

Coordination features across modalities such as bodily movements [60] and paralinguistic speech features [28] have been shown to be indicative of liking [60], attraction [60], and cohesion [28]. Here we test the hypothesis:

> *In an FCG, bodily coordination features (synchrony, convergence, mimicry, and causality) are positively correlated with PCQ.*

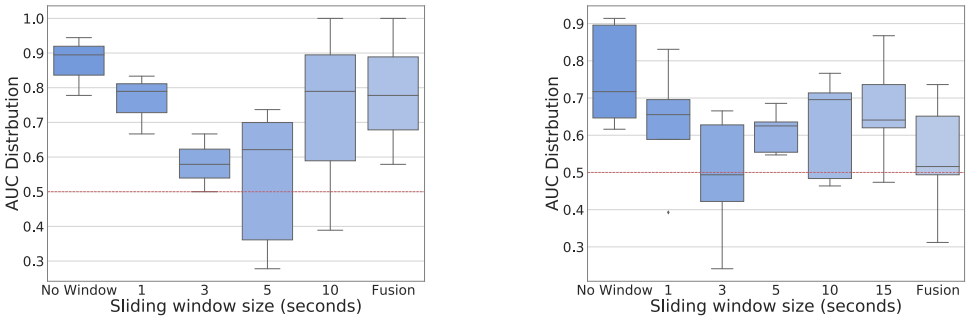
For the synchrony attributes, for both IndivPCQ and GroupPCQ we find that the *argmax* and *argmin* variants of lagged correlations are statistically significant attributes ($p = 0.003$). This suggests that the time taken to achieve maximum or minimum synchronous coordination has a significant effect on the conversation quality. We also note that for GroupPCQ, only correlation based features from the synchrony category were statistically significant, while other attribute sets (convergence and causality) were found to be statistically insignificant. For IndivPCQ, the *minimum* and *variance* of the convergence attributes were all statistically significant. This suggests that attributes capturing the least converging interacting pairs in a group are relevant to external observers. Moreover, we note that the minimum of the attributes are positively correlated, while the variance are negatively correlated. Further, the *maximum* and *minimum* of the lagged mimicry attributes were also statistically significant attributes. This suggests that pairs with high and low mimicry are relevant for estimating individual experience.

The Joint LASSO results indicate that several other feature sets also have a significant effect on IndivPCQ. Along with the *min*, *max*, *argmin*, and *argmax* attributes of the lagged correlation features, the non-lagged correlation were also significant. Moreover, the post-hoc rank correlation analysis associates different coefficient signs for some of the significant features. For example, lagged mimicry attributes are given negative β's by the rank correlation model but positive β's by LASSO. This suggests that there exists a non-linear monotonic relationships between these variables and IndivPCQ, causing the LASSO model to fail to explain this relationship, associating them with $\beta \approx 0$. One commonality between the two models is that both consider the *lagged* variant of mimicry features to be of more significance that the *lead* variant. For GroupPCQ, the LASSO and rank correlation analysis reveals that when jointly considered with other bodily coordination features, the lagged mimicry and convergence attributes are statistically significant.

8.6.2 Analysis of Feature Extraction and Fusion

Influence of Window Sizes

During data preprocessing we extract statistical and spectral features from the accelerometer data using the commonly used sliding window approach [58–60]. The choice of window-size influences a trade-off between noise-reduction and information loss. To



(a) GroupPCQ results

(b) IndivPCQ results

Figure 8.9: Results of the experiments on the predictive capabilities of different window-sizes.

understand the effect of this choice, we extract features using different window-sizes and evaluate the resulting change in the logistic regression model's predictive capability. The results are presented in Figure 8.9 for respective sliding window-sizes, along with the fusion of features from all the window-sizes, denoted as "*Fusion*".

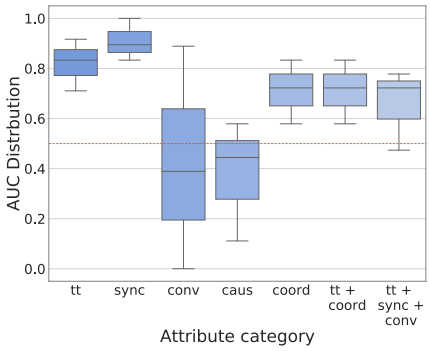
From Figure 8.9, we see that the best performing features are the ones where no sliding-window technique was used for both GroupPCQ and IndivPCQ. This suggests that the smoothing of accelerometer readings results in a loss of information which hurts model performance. The results might also indicate that bodily coordination between interacting pairs occur at finer temporal granularity, which can be captured directly without the sliding-window approach. The model with no sliding-window based features is capable of predicting GroupPCQ with a mean AUC of 0.85 ± 0.07 and IndivPCQ with a mean AUC of 0.76 ± 0.13. Also, noting here that using no sliding-window achieves the least standard deviation in AUC scores.

8

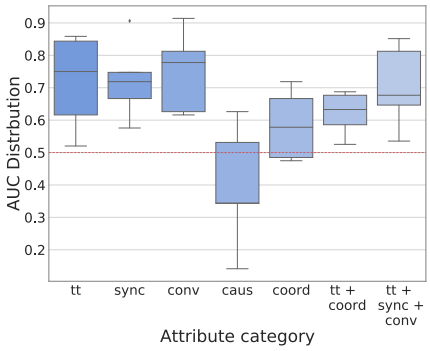
Influence of Fusing Attribute Categories

Here we study the influence of fusing different attribute categories on the performance of the logistic regression.

From the GroupPCQ results in Figure 8.10a, we see that the synchrony attributes (mean AUC of 0.89 ± 0.04) and turn-taking attributes (mean AUC of 0.81 ± 0.06), are the best performing attributes. In contrast to the IndivPCQ results in Figure 8.10b, the convergence attributes do not predict GroupPCQ well. Moreover, unlike for IndivPCQ, fusing turn-taking attributes with synchrony and convergence attributes does not improve GroupPCQ prediction, both in-terms of mean and variance AUC. From the IndivPCQ analysis, we see that convergence (mean AUC of 0.75 ± 0.12) and synchrony (mean AUC of 0.72 ± 0.12) based attributes perform well both by themselves and after feature-level fusion (mean AUC of 0.60 ± 0.10). We also observe that although turn-taking attributes are one of the best



(a) GroupPCQ results



(b) IndivPCQ results

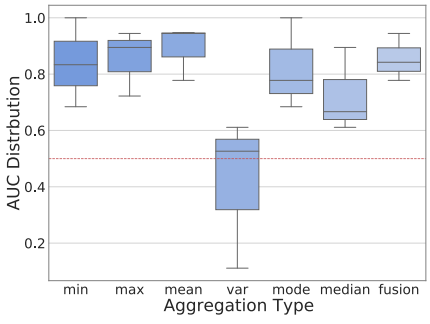
Figure 8.10: Predictive performance of different feature fusion approaches. Attribute category and indices as in Table 8.1—*tt*: Turn-taking (10-12), *sync*: Synchrony (1-4), *caus*: Causality (5-6), *conv*: Convergence (7-9), *coord*: Bodily Coordination (1-9).

performing feature sets by themselves (mean AUC of 0.72 ± 0.15), fusing them with bodily coordination attributes reduces the standard deviation of AUC, 0.70 ± 0.09. The results also suggest that synchrony and convergence attributes are best predictors of IndivPCQ, both individually and fused.

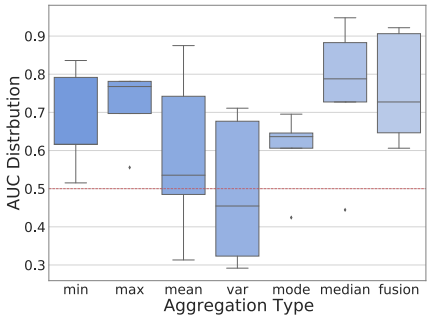
Influence of Feature Aggregators

The last step of our feature extraction procedure is to use aggregators to combine pairwise features into group-level features, or aggregate over pairs containing an individual for individual-level modeling, following previous works [5, 28, 30]. Here we study how different aggregators affect the predictive performance of the logistic regression model.

From Figure 8.11, we see that the *mean* aggregation of the features performs the best



(a) GroupPCQ results



(b) IndivPCQ results

Figure 8.11: Predictive performance of feature aggregators.

with a mean AUC of 0.89 ± 0.08. The mean is a skewed average. In contrast, for IndivPCQ the unskewed average, the *median*, is the most informative, with an AUC of 0.78 ± 0.17. This is in line with inferences drawn by Nanninga et al. [28] while studying cohesion in meetings. For both IndivPCQ and GroupPCQ, the *variance* aggregator performs worst.

8.7 Discussion and Conclusion

In this work, we have conceptualized, validated, and analyzed a perceived measure of conversation quality by unifying overlapping constructs that have so far been largely studied in isolation in literature. While our core motivation has been to gain insight into how people perceive the individual and group experiences of others, we do not claim that our proposed method measures, or is meant to be a third-party proxy for, the *one true experience* of the individual or group in the scene. On the contrary, we suggest that these perceptions are indicative of empathized gestalt impressions people draw of others' experience as it unfolds. We argue that such a perceived measure should complement other self-reported measures of experience to gain richer insight into how these differ and identify the contextual factors that influence the perceptions.

Third-party ratings are always prone to be influenced by biases that are heavily embedded in our cultures. We recommend users of this research to be mindful that third-party perceptions are not the same as self-reported measures. This fundamentally influences the system design process. The motivation for taking a third-party perspective is to enable a study of whether such perceptions have agreement, and whether samples with high agreement have common behavioral manifestations. To develop systems for inferring an individual's actual social experience, we advocate for a participant-in-the-loop strategy that allows for the measuring of the actual experience while being mindful of the participants' consent.

Inter-rater agreement and annotation drift are important aspects to consider while collecting annotations for behavioral data. Annotation drift is an issue when the annotator's mental model of the measured phenomenon changes over time while the phenomenon remains constant. Accounting for drift is crucial when the annotation is used as an attribute of the underlying phenomenon rather than as an attribute of the third-party observer. This is the case for annotating phenomena such as facial action units, where the goal of the annotation is to represent the configuration of a person's facial muscles rather than the annotator's perception of it, so a systematic drift over time or disagreement amongst annotators is undesirable.

For a perceived measure like the one we are proposing, the central phenomenon being studied is an onlooker's perception. So, every perception is inherently valid. This argument is based on our understanding that the measure requires some projection of one's own

experience onto the observed subjects when trying to empathize with their situation or take their perspective. Following the assumption that we construct narratives of other's behaviors, and that our appraisal of a situation is constructed based on our experiences, any drift occurring because of variations in one's experience can only provide (another) valid perspective on how the observed subject might be feeling. The same is true for variations in annotator agreement resulting from differences in perception of the annotators, either resulting from transient factors such as mood, or relatively stable factors such as personality and cultural background. For a perceived measure, we view all such perceptions as valid.

Designing the instrument to remove such variations would amount to artificially tampering with the phenomenon being measured. In our experiments we remove data with low inter-annotator agreement from the evaluation. However, this is because by design, the goal of the experiments is to gain insight into behavioral features that correlate with a high agreement on PCQ across raters. More broadly, we view the presence of low agreement on certain samples as a motivation for future work to explore more appropriate ways to embed subjectivity into the learning process when the goal is to train machine learning systems. Note that omitting the samples with low agreement from our experiments does not detract the validity of our measure. When the goal is to measure conversation quality as experienced by the individual or group in the scene, or even to use the third-party annotations as a proxy for the true experienced quality, we suggest treating the considerations of annotation drift and inter-rater agreement with care.

8.7.1 Limitations and Future Avenues

The data analyzed here was from spontaneous interactions in a single setting, that of mingling interactions following a speed-dating event. So, our findings pertaining to the individual features being indicative of PCQ ought to be interpreted within the scope of such a social context rather than being reflective of social behavior in all spontaneous interactions. As dedicated techniques for the non-invasive recording in-the-wild spontaneous interactions [74] continue to advance, it would be interesting to compare the effects of different social settings on the perception of PCQ using our proposed instrument.

Our operationalization of a conversing group follows the widely used framework F-formation [46]. However, recent evidence suggests that there might be multiple simultaneous conversations within a single F-formation containing more than four participants [72]. It would therefore also be interesting for future work to study PCQ within a single conversation floor rather than for the whole F-formation.

Finally, we have used three raters in this work to obtain our annotations. It would be useful for future works to use the proposed instrument to investigate systematic differences in perceptions of conversation quality across different cultures and demographics at scale.

Acknowledgments

This research was partially funded by the Netherlands Organization for Scientific Research (NWO) under the MINGLE project number 639.022.606. We thank Swathi Yogesh, Divya Suresh Babu, and Nakul Ramachandran for their time and patience in annotating the dataset, and Tiffany Matej Hrkalovic and Amelia Villegas Morcillo for the insightful discussions.

References

[1] A. Milek, E. A. Butler, A. M. Tackman, et al. "Eavesdropping on happiness" revisited: A pooled, multisample replication of the association between life satisfaction and observed daily conversation quantity and quality. *Psychological Science*, 29(9), 2018. doi: 10.1177/0956797618774252.

[2] M. R. Mehl, S. Vazire, S. E. Holleran, and C. S. Clark. Eavesdropping on happiness: Well-being is related to having less small talk and more substantive conversations. *Psychological Science*, 21(4), 2010. doi: 10.1177/0956797610362675.

[3] A. See, S. Roller, D. Kiela, and J. Weston. What makes a good conversation? how controllable attributes affect human judgments. In *Proc. of Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, June 2019. doi: 10.18653/v1/N19-1170.

[4] E. Dinan, V. Logacheva, V. Malykh, et al. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer, 2020.

[5] P. Müller, M. X. Huang, and A. Bulling. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In *Int., Conf., on Intelligent User Interfaces*, New York, USA, 2018. doi: 10.1145/3172944.3172969.

[6] N. Jaques, D. McDuff, Y. L. Kim, and R. Picard. Understanding and predicting bonding in conversations using thin slices of facial expressions and body language. *Lecture Notes in Comp., Science*, pages 64–74, 2016. doi: 10.1007/978-3-319-47665-0_6.

[7] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio. Detecting group interest-level in meetings. In *ICASSP*, Philadelphia, USA, 2005.

[8] C. Oertel, C. De Looze, S. Scherer, et al. Towards the automatic detection of involvement in conversation. In *Analysis of Verbal and Nonverbal Communication and Enactment.*, pages 163–170. Springer, 2011.

[9] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung. The MatchNMingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Trans., on Affective Computing*, 2018. doi: 10.1109/TAFFC.2018.2848914.

[10] C. Raman, J. Vargas-Quiros, S. Tan, et al. Conflab: A rich multimodal multisensor dataset of free-standing social interactions in-the-wild. *arXiv preprint arXiv:2205.05177*, 2022.

[11] J. L. Hagad, R. Legaspi, M. Numao, and M. Suarez. Predicting levels of rapport in dyadic

interactions through automatic detection of posture and posture congruence. In *IEEE Int., Conf., on Privacy, Security, Risk and Trust and Social Comp.*, pages 613–616, Oct 2011. doi: 10.1109/PASSAT/SocialCom.2011.143.

[12] R. Cuperman and W. Ickes. Big Five Predictors of Behavior and Perceptions in Initial Dyadic Interactions: Personality Similarity Helps Extraverts and Introverts, but Hurts "Disagreeables". 97(4):667–684, 2009. doi: 10.1037/a0015741.

[13] D. A. Northrup. *The problem of the self-report in survey research.* Institute for Social Research, York Univ., 1997.

[14] J. Garcia and A. R. Gustavson. The science of self-report. *APS Observer*, 10(1), 1997.

[15] L. J. R. Norman M. Bradburn and S. K. Shevell. Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys. In *New Series 1987*, pages 157–167. 1987.

[16] S. E. Lindley and A. F. Monk. Measuring social behaviour as an indicator of experience. *Behaviour & Information Technology*, 32:968–985, Oct. 2013. doi: 10.1080/0144929X.2011.582148.

[17] N. Raj Prabhu, C. Raman, and H. Hung. Defining and quantifying conversation quality in spontaneous interactions. In *Comp., Publication of the 2020 Int., Conf., on Multimodal Interaction*, pages 196–205, 2020.

[18] D. Reitter, J. D. Moore, and F. Keller. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proc., of the 28th Annual Conf., of the Cognitive Science Society*, pages 685–690, 2006.

[19] C. Oertel, S. Scherer, and N. Campbell. On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation. *INTERSPEECH*, pages 1541–1544, August, 2011.

[20] D. Wyatt, T. Choudhury, and H. Kautz. Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort. In *ICASSP*, Honolulu, USA, 2007.

[21] J. H. Antil. Conceptualization and operationalization of involvement. *Advances in Consumer Research*, 11(1), 1984.

[22] J. C.-y. Hsiao, W.-r. Jih, and J. Y.-j. Hsu. Recognizing continuous social engagement level in dyadic conversation by using turn-taking and speech emotion patterns. In *Workshop at Conf. on Artificial Intelligence*, pages 40–43, 2012.

[23] F. J. Bernieri, J. S. Gillis, J. M. Davis, and J. E. Grahe. Dyad rapport and the accuracy of its judgment across situations: A lens model analysis. *Journal of Personality and Social Psychology*, 71(1):110, 1996.

[24] A. O. Horvath and L. S. Greenberg. Development and validation of the working alliance inventory. *Journal of counseling psychology*, 36(2):223, 1989.

[25] C. Oertel and G. Salvi. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *Proc. of Int. Conf. on multimodal interaction*, pages 99–106, 2013.

[26] M. Casey-Campbell and M. Martens. Sticking it all together: A critical assessment of the group cohesion–performance literature. *International Journal of Management Reviews*, 11, 05 2009. doi: 10.1111/j.1468-2370.2008.00239.x.

[27] H. Hung and D. Gatica-Perez. Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behaviour. *IEEE Trans. on Multimedia*, pages 563–575, 2010. doi: 10.1109/TMM.2010.2055233.

[28] M. C. Nanninga, Y. Zhang, N. Lehmann-Willenbrock, Z. Szlávik, and H. Hung. Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry. In *Proc., of the 19th ACM Int., Conf., on Multimodal Interaction*, pages 206–215, 2017.

[29] Y. Zhang, J. Olenick, C.-H. Chang, S. W. J. Kozlowski, and H. Hung. The I in team: Mining personal social interaction routine with topic models from long-term team data. In *23rd Int. Conf. on Intelligent User Interfaces*, New York, USA, Mar. 2018. doi: 10.1145/3172944.3172997.

[30] Y. Zhang, F. Palo Alto Laboratory, U. Jeffrey Olenick, et al. TeamSense: Assessing Personal Affect and Group Cohesion in Small Teams through Dyadic Interaction and Behavior Analysis with Wearable Sensors. *Proc. of Interact. Mob. Wearable Ubiquitous Tech.*, 2018. doi: 10.1145/3264960.

[31] A. Cerekovic, O. Aran, and D. Gatica-Perez. How do you like your virtual agent?: Human-agent interaction experience through nonverbal features and personality traits. In *Human Behavior Understanding*, 2014.

[32] C. Edelsky. Who's got the floor? *Language in Society*, 10(3):383–421, 1981.

[33] J. Coates. Gossip revisited: Language in all-female groups. *Women in their speech communities*, 1989.

[34] M. Dunne and S. H. Ng. Simultaneous speech in small group conversation: All-together-now and one-at-a-time? *Journal of Language and Social Psychology*, 13(1):45–71, 1994.

[35] D. Tannen. *Conversational Style: Analyzing Talk among Friends.* Oxford University Press, 2005.

[36] A. F. Monk and D. J. Reed. Telephone conferences for fun: experimentation in people's homes. In *International Conference on Home-Oriented Informatics and Telematics*, pages 201–214. Springer, 2007.

[37] J. Carletta, S. Garrod, and H. Fraser-Krauss. Placement of authority and communication patterns in workplace groups: The consequences for innovation. *Small Group Research*, 29(5):531–559, 1998.

[38] O. Daly-Jones, A. Monk, and L. Watts. Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus. *International Journal of Human-Computer Studies*, 49(1):21–58, 1998.

[39] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

[40] N. Ambady and R. Rosenthal. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of personality and social psychology*, 64

(3):431, 1993.

[41] F. J. Bernieri, J. M. Davis, R. Rosenthal, and C. R. Knee. Interactional synchrony and rapport: Measuring synchrony in displays devoid of sound and facial affect. *Personality and social psychology bulletin*, 20(3):303–311, 1994.

[42] C. Raman, H. Hung, and M. Loog. Social processes: Self-supervised meta-learning over conversational groups for forecasting nonverbal social cues. *arXiv preprint arXiv:2107.13576*, 2021.

[43] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung. The MatchNMingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, 2018.

[44] C. Lai and G. Murray. Predicting group satisfaction in meeting discussions. *Proc. of Workshop on Modeling Cognitive Processes from Multimodal Data*, 2018. doi: 10.1145/3279810.3279840.

[45] S. W. J. Kozlowski and K. J. Klein. A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. *Multilevel theory, research and methods in organizations: Foundations, extensions, and new directions*, pages 3–90, 2000.

[46] A. Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[47] N. A. Murphy, J. A. Hall, M. Schmid Mast, et al. Reliability and validity of nonverbal thin slices in social interactions. *Personality and Social Psychology Bulletin*, 41(2):199–213, 2015.

[48] M. Z. Wang, K. Chen, and J. A. Hall. Predictive validity of thin slices of verbal and nonverbal behaviors: Comparison of slice lengths and rating methodologies. *Journal of Nonverbal Behavior*, 45:53–66, 2020.

[49] N. A. Murphy, J. A. Hall, M. A. Ruben, et al. Predictive validity of thin-slice nonverbal behavior from social interactions. *Personality and Social Psychology Bulletin*, 45:983–993, 2018.

[50] W. Brinkman. *Design of a Questionnaire Instrument*, pages 31–57. Nova Publishers, 2009.

[51] L. J. Cronbach and P. E. Meehl. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281, 1955.

[52] D. C. Funder and C. D. Sneed. Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of personality and social psychology*, 64(3):479, 1993.

[53] D. J. Mundfrom, D. G. Shaw, and T. L. Ke. Minimum sample size recommendations for conducting factor analyses. *International journal of testing*, 5(2):159–168, 2005.

[54] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Trans. on Audio, Speech, and Language Processing*, 17 (3), 2009.

[55] J. Cohen. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[56] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.

[57] M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[58] H. Hung, G. Englebienne, and J. Kools. Classifying social actions with a single accelerometer. In *Proc. of Int. Joint Conf. on Pervasive and Ubiquitous Comp.*, 2013.

[59] E. Gedik and H. Hung. Speaking status detection from body movements using transductive parameter transfer. In *Proc. of ACM Int. Joint Conf. on pervasive and ubiquitous computing*, pages 69–72, 2016.

[60] Ö. Kapcak, J. Vargas-Quiros, and H. Hung. Estimating romantic, social, and sexual attraction by quantifying bodily coordination using wearable sensors. In *2019 8th ACIIW*, pages 154–160. IEEE, 2019.

[61] E. Gedik and H. Hung. Detecting conversing groups using social dynamics from wearable acceleration: Group size awareness. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Dec. 2018. doi: 10.1145/3287041.

[62] E. Delaherche, M. Chetouani, A. Mahdhaoui, et al. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012. doi: 10.1109/T-AFFC.2012.12.

[63] J. Aldrich et al. Correlations genuine and spurious in pearson and yule. *Statistical science*, 10, 1995.

[64] D. C. Richardson and R. Dale. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive science*, 29 (6), 2005.

[65] J. Edlund, M. Heldner, and J. Hirschberg. Pause and gap length in face-to-face interaction. In *ISCA INTERSPEECH*, pages 2779–2782, Sept. 2009. doi: 10.21437/Interspeech.2009-710.

[66] J. Michalsky and H. Schoormann. Pitch convergence as an effect of perceived attractiveness and likability. In *INTERSPEECH*, pages 2253–2256, 2017.

[67] C. Lai, J. Carletta, and S. Renals. Modelling participant affect in meetings with turn-taking features. In *Proc. Workshop of Affective Social Speech Signals*, 2013.

[68] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33, 2010.

[69] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[70] E. Gedik and H. Hung. Detecting conversing groups using social dynamics from wearable acceleration: Group size awareness. *Proc. of Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4), 2018.

[71] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier, 1978.

[72] C. Raman and H. Hung. Towards automatic estimation of conversation floors within F-formations. *Int. Conf. on Affective Computing and Intelligent Interaction Workshops and Demos*, 2019. doi: 10.1109/ACIIW.2019.8925065.

[73] S. E. Lindley and A. F. Monk. Social enjoyment with electronic photograph displays: Awareness and control. *International Journal of Human-Computer Studies*, 66(8):587–604, 2008.

[74] C. Raman, S. Tan, and H. Hung. A modular approach for synchronized wireless multimodal multisensor data acquisition in highly dynamic social settings. In *Proc. of Int. Conf. on Multimedia*, 2020.

[75] P. Virtanen, R. Gommers, T. E. Oliphant, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[76] G. Varni, M. Avril, A. Usta, and M. Chetouani. Syncpy: a unified open-source analytic library for synchrony. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*, pages 41–47, 2015.

Supplementary Material

8.A PCQ Questionnaires

The questionnaire items below have been organized in terms of the different constituents of PCQ. The numbering before each questionnaire item indicates the ordering of the items in the original questionnaire. The original source from which the item was adapted is provided at the end of each question.

Instruction for the annotators: Use the set of questions below to annotate your perception of the individual and group's conversation quality, as seen in the video. First annotate the individual conversation quality (Section 8.A.1) for all the members of the group, and then annotate for group conversation quality (Section 8.A.2). Each interaction aspect in the below questionnaire should be rated using a five-point likert scale (Disagree strongly (1) to Agree strongly (5)). Read the questions carefully and observe the whole group carefully before annotating the video. You are allowed to re-watch the video again if required. For the group-level measures rate the group's behavior according to how you perceive them to behave *as a whole*. For ratings at both levels, try not to imagine how you would feel in their position, but focus on how they seem to feel based on their behaviour.

8.A.1 The Individual's Experience of Conversation Quality

Interpersonal Relationships

8 The individual was paying attention to the interaction throughout. [12]

9 The individual seemed to have gotten along with the group pretty well. [12][6]

Nature of Interaction

1 The individual looked like they had a smooth, natural, and relaxed interaction. [12]

2 The individual looked like they enjoyed the interaction. [12]

3 The individual's interaction seemed to be forced, awkward, and strained. [12]

4 The individual looked like they had a pleasant and an interesting interaction. [12]

5 The individual appeared uncomfortable during the interaction. [12]

10 The individual appeared self-conscious during the interaction. [12]

Equal Opportunity

6 The individual attempted to take the lead in the conversation. [6][31]

7 The individual looked like they experienced a free-for-all interaction. [16]

8.A.2 The Group's Conversation Quality

Interpersonal Relationships

4 The group members seemed to have accepted and respected each other in the interaction. [12]

7 The group members seemed to have gotten along with each other pretty well. [12][6]
8 The group members were paying attention to their partners throughout the interaction. [12]
9 The group members attempted to get "in-sync" with their partners. [12][6]
10 The group members used their partner's behavior as a guide for their own behavior. [12][6]

Nature of Interaction

1 The interaction within the group seemed smooth, natural and relaxed. [12]
2 The group members seemed to have enjoyed the interaction. [12]
3 The interaction within the group seemed forced, awkward, and strained. [12]

Equal Opportunity

5 The group members seemed to have received equal opportunity to participate freely in the interaction. [16]
6 The interaction involved equal participation from all group members. [16]

8.B Feature Extraction Details

8.B.1 Synchrony

Correlation

As a measure of correlation, in this research, we use the *Pearson correlation coefficient* (using the *pearsonr* method available in the scipy package [75]) to measure the correlation. The pearson correlation coefficient is calculated as follows,

$$\rho_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sigma(X)\sigma(Y)} \tag{8.1}$$

where, x and y are the preprocessed accelerometer data of length N from person X and Y, x_i and $y + i$ are the data values of x and y respectively at time-step i, μ_x and μ_y are the means of x and y respectively, and σ_y and σ_y are the standard-deviations of x and y respectively.

Time-lagged Correlation

The time-lagged correlation is computed using pearson correlation coefficients at different time lags, as follows,

$$\rho_{XY} = \frac{\sum_{i=1}^{N-\tau}(x_i - \mu_x)(y_{i+\tau} - \mu_y)}{\sigma(X)\sigma(Y)} \tag{8.2}$$

where, X and Y are the preprocessed accelerometer data of length N from person X and Y, x_i and $y + i$ are the data values of x and y respectively at time-step i, μ_x and μ_y are the means of X and Y respectively, and σ_y and σ_y are the standard-deviations of X and Y respectively. More importantly, the variable τ denotes the time-lag, that is, the positive time-lag in terms of time steps between X and Y.

Mutual Information

We use *Mutual Information* to capture the degree of dependence of signal values between two interlocutors. It is calculated as follows,

$$MI(X,Y) = \frac{H(X) + H(Y) - H(X,Y)}{\sqrt{H(X)H(Y)}} \tag{8.3}$$

where H(X) and H(Y) denotes the entropy of preprocessed accelerometer data of person X and person Y, and H(X, Y) represents the joint entropy of these preprocessed accelerometer data X and Y. To calculate this feature, we use the Mutual Information calculator available in the *SyncPy* package [76].

Mimicry

As the *Mimicry* measure, we use the similar technique as Nanninga et al. [28]. The mimicry metric used in [28] was originally extracted from paralinguistic signals, in our case, we extract these features from the preprocessed accelerometer data. The extraction technique is depicted in Figure 8.12.

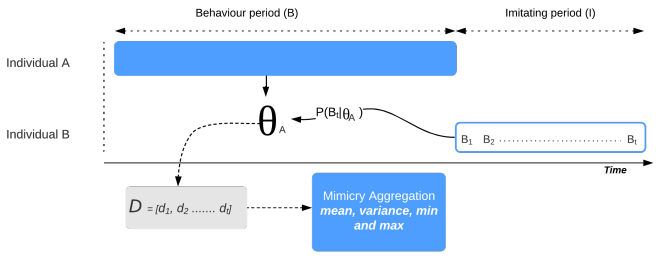


Figure 8.12: The illustration shows the extraction of Lagged Mimicry for Individual B and Lead Mimicry for Individual A. The θ is the learnt model from A's behaviour period and $B_1, B_2 ..... B_t$ are the data samples at each time stamp of Individual B. A distance vector D is computed with respected to the probability values $P(B_t|\theta_A)$, which is later used to extract aggregate based mimicry features.

8.B.2 Causality

Coherence

In studying social signal processing, Richardson and Dale [64] have used coherence based methods to study discourse comprehension of speakers and listeners. The coherence between two signals X and Y can be measured as follows,

$$C_{XY}(f) = \frac{|G_{XY}(f)|^2}{G_{XX}(f)G_{YY}(f)} \tag{8.4}$$

where, X and Y are the preprocessed accelerometer data of person X and Y, $G_{XY}(f)$ corresponds to the cross-spectral density of a signal and $G_{XX}(f)$ and $G_{YY}(f)$ correspond to the auto-spectral density of signals X and Y respectively. Values of coherence will always satisfy the property: $0 \leq C_{XY}(f) \leq 1$. To calculate this feature, we use the Coherence calculator available in the *SyncPy* package [76].

Causality

The Granger's causality test is a statistical test which, similar to coherence, captures the causality of one signal over another but in a different manner. This particular measures capture the causality by estimating whether one signal is useful in forecasting the other signal. In particular, let $X(t) \in \mathbb{R}^{d\times 1}$ for $t = 1, \ldots, T$ be a d-dimensional multivariate signal. Granger causality is performed by fitting a VAR model with L time-lags as follows,

$$X(t) = \sum_{\tau=1}^{L} A_\tau X(t-\tau) + \varepsilon(t) \tag{8.5}$$

where $\varepsilon(t)$ is a white Gaussian random vector, and A_τ is a matrix for every τ. A signal X_i is called a granger cause of another time series X_j, if at least one of the elements $A_\tau(j, i)$ for $\tau = 1, \ldots, L$ is significantly larger than zero (in absolute value). In other words, an *f-test* is performed on the Ordinary Least Squares (OLD) model with the optimal lag (estimated using the BIC criterion), resulting in a f-value and a p-value which is open for interpretation. For this research, we use the Granger Causality calculator available in the *SyncPy* package [76].

8.B.3 Convergence

Symmetric Convergence

The symmetric convergence captures the decrease or increase in similarity between two body movements along time, *without any lag* between the two signals. The extraction technique is depicted in Figure 8.13.

Asymmetric Convergence

The asymmetric convergence captures the decrease or increase in similarity between two body movements along time, *with a time-lag* between the two signals. The extraction technique is depicted in Figure 8.14.

Global Convergence

Global convergence captures the change in similarity between two body movements, specifically between its initial time-segments and its later time-segments. The extraction technique is depicted in Figure 8.15.

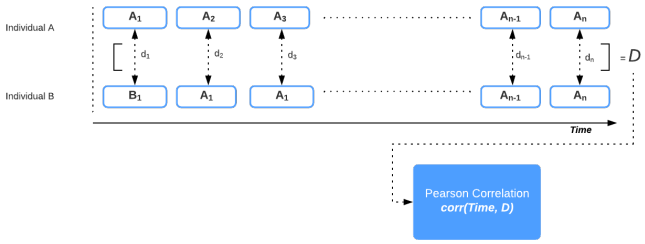


Figure 8.13: An illustration of symmetric convergence extraction between interacting partners A and B. $A_1, A_2....A_n$ and $B_1, B_2....B_n$ vectors represent the accelerometer readings from A and B, respectively. A distance vector D is calculated using squared distance between the A and B's data samples. Finally, D is used to computer the correlation with *time*, to capture the evolving similarity.

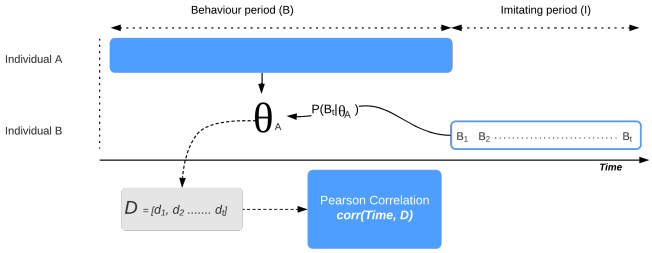


Figure 8.14: The illustration shows the extraction of asymmetric convergence between interacting partners Individual A and B. The θ is the learnt model from A's behaviour period and $B_1, B_2.....B_t$ are the data samples at each time stamp of Individual B. A distance vector D is computed with respected to the probability values $P(B_t|\theta_A)$, which is later used to computer the correlation with *time*, to capture the evolving similarity.

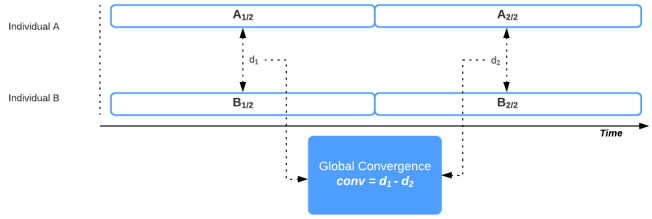


Figure 8.15: Illustration of global convergence between interacting partners A and B. Both A and B's accelerometer channels are split into two halves, $(A_{1/2})$, $(A_{2/2})$, $(B_{1/2})$ and $(B_{2/2})$, and squared distances (d_1, d_2) are computed with the respective halves. Finally, global convergence is the difference between the squared distances.

8.C Class Imbalance Distribution

Table 8.3: Class Distribution between high and low PCQ samples, after Kappa and label thresholds.

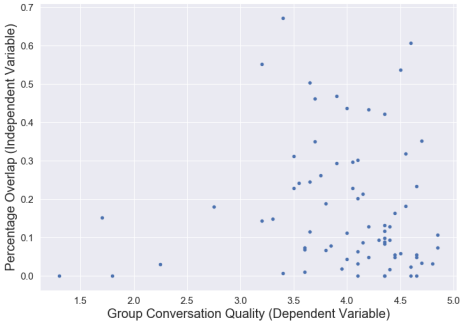
(a) For GroupPCQ.

Low GroupCQ	High GroupCQ
3	55

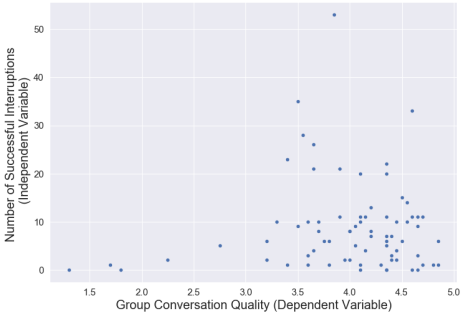
(b) For IndivPCQ.

Low IndivCQ	High IndivCQ
16	163

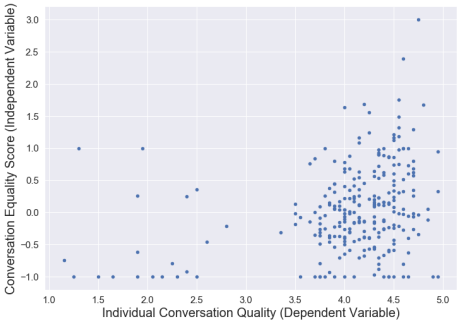
8.D Additional Figures



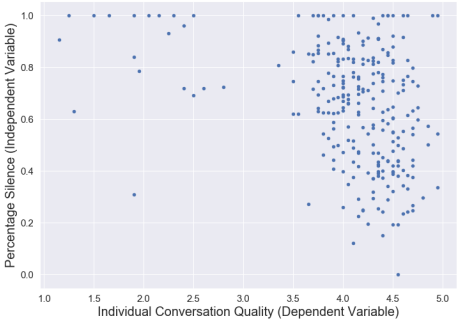
(a) Scatter plot revealing the relationship between Percentage of Overlap (Independent Variable) and the Group Conversation Quality (Dependent Variable).



(b) Scatter plot revealing the relationship between Number of Successful Interruptions (Independent Variable) and the Group Conversation Quality (Dependent Variable).



(c) Scatter plot revealing the relationship between Conversation Equality (Independent Variable) and the Individual Conversation Quality (Dependent Variable).



(d) Scatter plot revealing the relationship between Percentage of Silence (Independent Variable) and the Individual Conversation Quality (Dependent Variable).

Figure 8.16: Scatter plots with respect to few Independent Variables and the Dependent Variables of Conversation Quality. The scatter plots are a qualitative analysis of the independent variable's variance (σ^2) conditioned to the dependent variable of Conversation Quality, and thus examine the Exogeneity and Homoscedasticity of the dataset.

IV

Discussion

9

Discussion

One of the things Ford Prefect had always found hardest to understand about humans was their habit of continually stating and repeating the very very obvious.

— Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

An implicit goal of this Thesis has been to advance Artificial Social Intelligence (ASI) research in the wild along broad research themes and avenues. The preceding chapters have consequently spanned the themes of acquiring, modeling, and perceiving social human behavior. In the course of performing this research we have identified specific challenges and opportunities for taking computational social behavior research into natural real-world settings. Here we discuss the contributions of this Thesis towards this goal, the limitations of this work, and potential ways forward for ASI research in the wild.

9.1 Curating Social Behavior Datasets

A core challenge in studying real-world social behavior is the lack of representative data. In **Chapter 2** we addressed this challenge directly, proposing a replicable data collection concept called ConfLab viewing conferences as living labs. This enabled us to design a real-world interaction setup involving a diverse mix of seniority, acquaintanceship, and motivations for mingling. In doing so, we made choices and proposed technical innovations to reduce the cost, effort, and time in collecting such datasets, and maximize data fidelity while upholding ethical best practices. These innovations included the modular synchronization-at-acquisition method described in **Chapter 3** which achieves latency tolerances suitable for studying phenomena such as mimicry and synchrony without the need for post hoc data synchronization strategies. We also proposed the Covfee annotation tool to enable continuous annotation, capturing gestalt impressions of raters and lowering annotation time compared to traditional methods [1]. Finally, we developed the Midge wearable sensor in a noninvasive conference badge form factor improving upon previous wearable sensors by including a full 9-axis inertial measurement unit and the on-board ability to switch between audio recording frequencies to preserve participant privacy.

While these technical innovations are an essential step in sensing and annotating in-the-wild behavior, we have only provided the first iteration of ConfLab recording at a major international conference. Even this first iteration required significant cooperation and support from conference organizers, not to mention the logistical challenges arising from recording in a different country while ensuring compliance with ethical best practices established by multiple organizations and nations. However, in order to study the effect of short time-scale embodied behaviors on long term relationships in real social networks, multiple iterations of ConfLab are required. Our hope here is that the *data for the community by the community* ethos of ConfLab will find wider traction, allowing for amortizing the cost and effort of these data collection efforts across multiple research groups. In the meanwhile, researchers must be aware that the insights resulting from data from this first iteration may not generalize to other interaction contexts and the general population.

9.2 Data Efficient and Adaptive Modeling of Social Behavior

Chapter 5 addresses the limited-data challenge from a modeling perspective: if data is scarce, can we develop machine learning methodologies that use the available data efficiently to model social behavior? Here we formalized the Social Cue Forecasting task to learn representations of low-level behavioral cues in a self-supervised manner. The idea is to regress future social cues from the same preceding cues in a bottom-up manner, in contrast to traditional top-down approaches that predicted semantic behaviors or high-order social signals from cues. This enables learning general representations of behavior from the entire available data. Specifically, from social science insights we motivated the need to model a distribution over cues and jointly model future cues for all participants to account for behavioral interdependence. Beyond this task formulation that advocates for utilizing all available low-level cue data, we also proposed the Social Process models. By viewing conversing groups as meta-learning tasks, we showed how models can adapt to the unique behavioral coordination of unseen groups at evaluation. Specifically, we modeled the low-level dynamics of group behavior as stochastic processes, learning joint latent representations for all participants in the group. Crucially, unlike previous work, the proposed method is also invariant to group sizes and participant order.

The overarching motivation of this line of research is to develop techniques that can adapt to both individuals and groups from a few observations. The Social Process models incorporate the interdependence between conversation partners into a single latent variable. Consequently, a few natural questions arise regarding learning latent representations of behavior directly from data. What latent factors uniquely describe a group? What latent factors unique to individuals translate across the groups in which they participate? Here, one possibility for future work is to induce a hierarchical structure over the latent space, to model the interplay between individual and group latent variables. Doing so would also address a limitation of the work in Chapter 5: so far, we have ignored inter-group dynamics. In a complex conversational scene, social influence from outside a single group might motivate individuals to leave and interact with different partners. Learning the structure of the latent space for an entire scene, allowing for interplay between the individual and group latent factors, constitutes a promising direction for this line of research. Of course, researchers should be cognizant that learning structure directly from data with reduced inductive prior information generally requires more data. Another aspect not investigated in this Thesis is how the latent representations from these models can be utilized in downstream tasks. The representations have been trained to predict future low-level behavior while incorporating participant interdependence. Consequently, research into anticipating specific phenomena such as interaction termination or turn changes can

benefit from fine-tuning them using specific supervisory labels, or developing techniques for imposing further structure over the latent space in the study of these phenomena.

In modeling low-level behavioral cues, we emphasize the choice of feature representations. In our experiments in Chapter 5 we represented pose using quaternions. Beyond the favorable properties of quaternions discussed in the chapter, doing so also allowed us to have a uniform feature representation across the MatchNMingle and Haggling datasets that contain a different number of keypoints for every individual and different camera perspectives. Here, when representing horizontal rotations, two dimensions of the quaternions were perpetually zero throughout the data. In performing the experiments we discovered that such constant features can pose problems when optimizing the evidence lower bound (ELBO). Specifically, the models we experimented with maximized log-likelihood by making the variance over these features exceedingly small, often at the cost of learning a worse mean. Subsequent experiments with alternate representations such as keypoints did not suffer from such issues. Consequently, we advise that special care is taken in choosing feature representations and ensuring that the training procedure of such generative models does not suffer from well-studied issues such as mode-collapse [2].

9.3 Synthesizing Social Human Behavior

Beyond curating additional datasets and developing efficient modeling techniques, a promising new approach to dealing with limited representative real data is to synthesize it. **Chapter 4** takes the first step in this direction. Given that synthesizing rich and expressive multimodal behaviors remains a long-term goal, we first explore the more constrained task of synthesizing faces. Specifically, we find that boosting the realism of synthetic faces—with dynamic expression-based wrinkles in this case—helps in achieving performance on downstream computer vision tasks comparable to that using real data.

The promise of synthetic data lies in the control it affords in addressing the biases that exist in real data. Such biases can span lighting and environmental conditions as well as factors surrounding appearance, clothing, and cultural representation. However, synthesizing faces only scratches the surface of the possibilities of synthetic behavioral data. Generating multimodal behaviors inherently suffers from a chicken-egg problem: synthetic data is meant to address the scarcity of real data, but requires real data to learn from. In the absence of adequate representative data across interaction settings and cultures, one must again turn to existing insights from social theory to generate believable and varied synthetic cues. Another opportunity in this space is to generalize to groups beyond dyads. Most existing works in synthesizing nonverbal behavior have focused on nonsocial [3, 4] or dyadic settings [5–7]. Here, our work in **Chapter 5** on group-size agnostic modeling of low-level behavior is applicable in synthesizing socially-aware cues for multiple participants,

thereby generalizing beyond dyadic settings. Nevertheless, several challenges exist in this research space. How do we evaluate synthesized behaviors? How do synthesized cues relate to the outcomes an artificial agent might desire to achieve in an interaction? While a detailed discussion of these challenges is beyond the scope of this discussion, the research space of synthetic behavioral data affords rich research questions towards advancing ASI.

9.4 Ethics and Privacy: Behavior as Biometrics?

In **Chapter 2** we have discussed at length the trade-offs involved in improving data fidelity and concerns surrounding participant privacy and ethical considerations. Specifically, our participatory design principles have followed an agentist rather than structurist approach (see Section 2.7). The goal here is to enable individualized measurements of social behavior, in contrast to structural analyses commonplace in network sciences. Despite the aforementioned choices in protecting the sensitive visual and verbal information of participants, moving towards individualized measurements presents increased ethical concerns.

One consideration is the potential of nonverbal social behavior as biometric information. Biometric technology concerns the use of physiological and behavioral characteristics of individuals. While the first generation of biometrics focused on physiological individual identifiers, the second shifted focus to behaviors [8]. Schumacher [9] characterized this shift as moving from *who you are* to *how you are*. Behavioral features including gait, stride, lip movement, speech, and blinking have been found to provide sufficiently accurate identity verification [10]. Moving across these generations has also accompanied a shift in purpose and applications of biometrics, from security to applications in commercial and civil technology such as assessing student engagement in classrooms [8].

The advancement of behavioral biometrics and its introduction in daily life poses some problematic ethical concerns and privacy risks. If anonymity cannot be preserved, numerous types of privacy are violated. While informational privacy constituted the primary early concern [11, 12], the evolution of biometric technology accounts for seven types of privacy including the privacy of thoughts and feelings [13]. Obtaining informed consent can then be mired in power imbalances as passive sensing technology captures seemingly benign behavioral features from unaware subjects [14]. Moreover, technology trained on seemingly privacy-preserving modalities can acquire biases that can facilitate discriminatory decision-making while providing the illusion of objectivity [15].

Furthermore, ethical considerations also necessitate paying attention to legal and cultural factors. While ConfLab's data collection setup was compliant with the General Data Protection Regulation (GDPR) in EU Law, different national legal environments allow for different degrees of privacy when collecting data. For instance, China's Personal Information Privacy Law (PIPL) can be even stricter than GDPR in some regards. Meanwhile,

as privacy can be considered by some as a value rather than a right [8], attitudes towards privacy and ethics can also be influenced by cultural norms and values.

The discussion surrounding ethical considerations against a backdrop of ever-advancing AI technology is complex, and there are no direct and easy solutions to the unique ensuing concerns. It would therefore be beneficial for future research in the realm of ASI to give the matter greater consideration, especially when dealing with in-the-wild data.

9.5 Meta Discussion: Bringing Disciplines Closer - A Practitioner's Perspective

9.5.1 The Disciplinary Spectrum

The central motivation guiding the broader conception of ASI (**Chapter 1**) was emphasizing a bidirectional reciprocity between AI and the social science disciplines. If the field of ASI is so inherently interdisciplinary, where does the divide between disciplines arise? At one end of the spectrum are computer scientists. Being typically unversed in social theory, computer scientists often resort to explaining social phenomena using natural science theories including physical forces [16] and evolution [17] (also see [18, 19]), which may often be insufficiently expressive. This lack of social literacy bears the risk of devolving into a pseudoscience, where a combination of misappropriated models and *reading the tea leaves* [20] can lead to unsubstantiated social scientific claims or *folk theory* [21]. Similarly, at the other extreme of the spectrum are social scientists, who are typically unversed in advanced computational techniques such as machine learning or neural networks [22–24]. This lack of technical literacy can lead to unsubstantiated fears about the social implications of AI or mistrust of AI systems on the one hand [19, 24], as well as unrealistic expectations about the capabilities of such purportedly intelligent systems on the other. Between these extremes lie the interdisciplinary fields. Here, in the pursuit of being versed in multiple rapidly evolving disciplines, one runs the risk of knowing none.

The cumulative effect of these differences has been a largely unidirectional flow of knowledge between the disciplines. Amidst the growing incorporation of AI methods in sociology almost two decades ago, Chai [18] observed a lack of export of general social theories into AI. More importantly, within sociology itself, the cookie-cutter application of AI methods often accompanied an ignorance of domain knowledge: "for the most part, social simulations within sociology, rather than drawing on general social theory for their assumptions, have seemed to largely follow existing approaches from AI". In the present landscape, the exponential progress of deep learning research has catalyzed a similar trend. Large off-the-shelf architectures are often applied to social behavior data without adapting them using domain insights, or catering to domain challenges such as limited data. While

the bulk of the contemporary focus lies with data analysis and modeling, comparatively less is being done towards bottom-up frameworks for assisting social theory building. This points to a *producer-consumer* model of research between AI and the less computational-oriented disciplines interested in social phenomena. Even so, it is perhaps impractical to expect the common practicing researcher—grappling with the rapidly growing demands on their time and an even more rapidly growing pile of must-read literature—to keep abreast methods from multiple disciplines. Beyond formal education, our knowledge is acquired implicitly by reading articles and textbooks, whose authors may also not have given interdisciplinary considerations much thought. The pursuit of science still remains a human process, and is as such guided by methodological and cultural norms familiar to researchers within their research bubbles.

9.5.2 How then, can we bridge the divide?

One avenue is of course, formal education. Almost three decades ago, Bainbridge et al. [24] had already noted the need for literacy in interdisciplinary techniques, albeit on the sociology side: "Current graduate training does not prepare students to take advantage of ASI. Although some probability theory can be useful, hardly any of the material taught in statistics courses is relevant to the computer techniques described here." I posit that a similar training in social theory is warranted on the computer science side for researchers working on problems with social implications. In this respect, I believe things are moving in the right direction, with interdisciplinary programs of study becoming more commonplace. The second avenue is to interact more with researchers from other disciplines. Here, interdisciplinary communities, workshops, and conferences such as ICMI (https://icmi.acm.org/), ACII (https://acii-conf.net/) provide fertile breeding grounds for the cross-pollination of ideas. The practical risk here is that the greater perceived prestige associated with the more mono-disciplinary venues, combined with an academic reward system that largely prioritizes perceived prestige, results in a dichotomy: should researchers prioritize meaningful engagement that is more common in smaller interdisciplinary communities at the cost of publishing at prestigious venues? The dichotomy bears more weight for early-career researchers who are yet to establish themselves.

Yet, at a larger scale, these avenues might only be available to a select minority. For a sizable majority, formal education and being embedded in an environment with access to researchers from other disciplines are privileges. Nevertheless, information continues to become more accessible. For the reader who has chanced upon this Thesis, without any claims of being definitive or indeed prescriptive, I propose some rules of thumbs to use in their own work towards bridging the disciplinary divide.

Grounding in Domain Literature and Challenges

The machine learning practitioner might find utility in paying attention to the *formulation of the task* at hand, and whether it reflects the nature of the underlying phenomenon being modeled. For instance, in **Chapter 5**, the formalization of Social Cue Forecasting considered domain knowledge to argue for predicting a distribution over futures rather than the common practice of predicting a single future at a time. Moreover, the formalization also makes the case for jointly forecasting futures for all participants in an interaction given evidence for the interdependency between partner behaviors.

When it comes to designing *machine learning methodology*, practitioners might benefit from considering how their architectures or methods reflect domain knowledge. While expressing social theories mathematically is not always straightforward, it is an important matter to ponder. The Social Process models proposed in **Chapter 5** treat learning the unique adaptation of behaviors within a conversing group as a meta-learning task. In doing so, the models can generalize to unseen groups at test in a data-efficient manner.

Being Aware of Implicit Assumptions

Practitioners from all disciplines related to ASI would benefit from ensuring that the assumptions of the systems they are using matches their own. Within *machine learning and the computational fields*, this pertains to being aware of the datasets pretrained models were trained on and the possible biases they might have acquired. More straightforward, it also pertains to understanding what basic machine learning architectures such as convolutional or recurrent networks are designed to model. Within the *applied disciplines*, it is worthwhile to remember that modeling tasks often require making simplifying assumptions or approximations. Being aware of what these are may help researchers in revealing insights about the nature of social phenomena. For instance, **Chapter 7** established evidence to challenge the implicit assumption of *one conversation per conversing group* that prior computer vision works had made for the task of detecting conversing groups in scenes. It might be important to note that given the implicit nature of such methodological assumptions, one might need to draw upon one's own social experiences to identify and challenge them.

Beyond methodological assumptions, it is crucial to double check assumptions about the data. It is often easy to altogether overlook issues in the data. One source of issues could be the sensing and capture setup. For instance, the modular synchronization-at-acquisition solution presented in **Chapter 3** was aimed at ensuring that the latency in multimodal data supports fine-grained temporal analysis of social phenomena. Beyond synchronization, sensor calibration is another consideration to be aware of. Another source of date-related issues is annotation. Here, it is important to be aware of the annotation and ground-truthing procedure. Are there validity or reliability concerns? What was the demographics of the annotators? Would they have introduced biases? Keypoint

annotations, for instance, are usually obtained by annotating single frames separated in time, and interpolating for the interim frames. The lack of motion in per-frame annotations may introduce artificial artifacts across frames. This was one of the motivations for designing the continuous annotation framework Covfee [1] for annotating the ConfLab dataset presented in **Chapter 2**.

Supporting Research on the Other Side of the Fence

The third aspect to consider is how one's proposed work might benefit researchers from other domains, who might speak very different professional languages. One way is to help situate readers from different disciplines using literature and parallel perspectives familiar to them. Of course, this might require writing under the somewhat bold assumption that researchers from outside one's discipline would read the proposed work. It might also require convincing reviewers of the relevance of referring to broader literature.

Taking inspiration from the work of Nelson [25] and Bamman et al. [26], Radford and Joseph [27] aggregate two potential blueprints for how machine learning can aid in revising social theory. One way is to use known theories to hypothesize about what empirical results might look like, and to provide alternative hypotheses for what results might look like of a new or revised theory was instead true. Another is to build a machine learning model that matches a theoretical model, and then demonstrate how adding components inspired by new or revised theory improve model performance. In both cases, the new theory is always expected to originate from the researcher. Moreover, the blueprints require existing theories for the phenomenon being studied.

Combining **Chapters 5 and 6**, this Thesis establishes a first step toward a third blueprint: the automated generation of hypotheses for evaluation. We do so by proposing a method to identify patterns of data salient to a model trained in a bottom-up manner, thereby viewing the model as a human observer. Here, our experiment on real-world data illustrated that timesteps at which an individual rotated away from their interaction partners were salient for a Social Process model in predicting their group-leaving behavior in the future. In this case, this observation corroborates evidence in social science literature about behaviors predicting interaction termination. Nevertheless, we argue that such data-driven insights are strictly hypotheses obtained from real-world data which need subsequent investigation, and emphasize a research methodology that involves a domain expert in the loop. Here, demonstrating intended uses and limitations of proposed the framework is meant to situate researchers in applied domains. The present work stops at identifying timesteps rather than features that are salient for a given model; nevertheless the broader goal of automatically obtaining data-driven domain hypotheses constitutes an open research direction.

On the social sciences side, establishing theories pertaining commonly made assumptions in machine learning can influence the development of new modeling techniques. For

instance, the results from **Chapter 7** may motivate researchers interesting in detecting conversing groups to reconsider how the notion of a group is operationalized. Another way for the social sciences to support the machine learning research into ASI is to provide quantitative measures for subjective constructs so that artificial agents may perceive them. In this regard, **Chapter 8** argues for a *perceived* measure of conversation quality in light of the fact that the true measure of the quality of a conversation experienced by an individual is never known, even to humans. In interactions, we use a Theory of Mind to evaluate our partners' experience of conversation quality, so it would benefit artificial agents to similarly have a conception of the perceived conversation quality to conduct social interactions in a manner that exhibits social awareness.

References

[1] J. Vargas Quiros, S. Tan, C. Raman, L. Cabrera-Quiros, and H. Hung. Covfee: an extensible web framework for continuous-time annotation of human behavior. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, volume 173 of *PMLR*. PMLR, 2022.

[2] H. Fu, C. Li, X. Liu, et al. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. In *Proceedings of NAACL-HLT*, pages 240–250, 2019.

[3] C. Ahuja, D. W. Lee, Y. I. Nakano, and L.-P. Morency. Style Transfer for Co-Speech Gesture Animation: A Multi-Speaker Conditional-Mixture Approach. *arXiv:2007.12553 [cs]*, July 2020.

[4] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, and M. Neff. A comprehensive review of data-driven co-speech gesture generation. *arXiv preprint arXiv:2301.05339*, 2023.

[5] C. Ahuja, S. Ma, L.-P. Morency, and Y. Sheikh. To React or not to React: End-to-End Visual Pose Forecasting for Personalized Avatar during Dyadic Conversations. *arXiv:1910.02181 [cs]*, 2019.

[6] N. T. V. Tuyen and O. Celiktutan. Context-aware human behaviour forecasting in dyadic interactions. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 88–106. PMLR, 2022.

[7] Y. Yoon, P. Wolfert, T. Kucherenko, et al. The GENEA challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, pages 736–747, New York, NY, USA, 2022. ACM.

[8] A. North-Samardzic. Biometric technology and ethics: Beyond security applications. *Journal of Business Ethics*, 167(3):433–450, 2020.

[9] G. Schumacher. *Behavioural Biometrics: Emerging Trends and Ethical Risks*, pages 215–227. Springer Netherlands, Dordrecht, 2012. doi: 10.1007/978-94-007-3892-8_10.

[10] R. V. Yampolskiy and V. Govindaraju. Behavioural biometrics: a survey and classification. *International Journal of Biometrics*, 1(1):81–113, 2008.

[11] A. Cavoukian, M. Chibba, and A. Stoianov. Advances in biometric encryption: Taking privacy by design from academic research to deployment. *Review of Policy Research*, 29(1):37–61, 2012.

doi: 10.1111/j.1541-1338.2011.00537.x.

[12] G. J. Smith, M. San Roque, H. Westcott, and P. Marks. -surveillance texts and textualism: Truthtelling and trustmaking in an uncertain world. *Surveillance & Society*, 11(3):215–221, 2013.

[13] R. L. Finn, D. Wright, and M. Friedewald. *Seven Types of Privacy*, pages 3–32. Springer Netherlands, Dordrecht, 2013. doi: 10.1007/978-94-007-5170-5_1.

[14] A. Norval and E. Prasopoulou. Public faces? a critical exploration of the diffusion of face recognition technologies in online social networks. *New Media & Society*, 19(4):637–654, 2017.

[15] K. Martin. Ethical implications and accountability of algorithms. *Journal of Business Ethics*, 160: 835–850, 2019.

[16] D. Helbing and P. Molnar. Social Force Model for Pedestrian Dynamics. *Physical Review E*, 51 (5):4282–4286, May 1995. doi: 10.1103/PhysRevE.51.4282.

[17] M. Mastrangeli, M. Schmidt, and L. Lacasa. The roundtable: An abstract model of conversation dynamics. *arXiv:1010.2943 [physics]*, Oct. 2010.

[18] S. Chai. Artificial intelligence and social theory: A one way street. *Perspectives (Gerontological Nursing Association (Canada))*, 27(4):11–12, 2004.

[19] J. Mökander and R. Schroeder. AI and social theory. *AI & SOCIETY*, 37(4):1337–1351, Dec. 2022. doi: 10.1007/s00146-021-01222-z.

[20] J. Chang, S. Gerrish, C. Wang, J. Boyd-graber, and D. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[21] R. G. d'Andrade. *The Development of Cognitive Anthropology*. Cambridge University Press, 1995.

[22] M. Molina and F. Garip. Machine Learning for Sociology. *Annual Review of Sociology*, 45(1): 27–45, 2019. doi: 10.1146/annurev-soc-073117-041106.

[23] K. M. Carley. Artificial intelligence within sociology. *Sociological Methods & Research*, 1996.

[24] W. S. Bainbridge, E. E. Brent, K. M. Carley, et al. Artificial Social Intelligence. *Annual Review of Sociology*, 20:407–436, 1994.

[25] L. K. Nelson. Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*, 49(1):3–42, Feb. 2020. ISSN 0049-1241. doi: 10.1177/0049124117729703.

[26] D. Bamman, T. Underwood, and N. A. Smith. A Bayesian Mixed Effects Model of Literary Character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379. Association for Computational Linguistics, 2014. doi: 10.3115/v1/P14-1035.

[27] J. Radford and K. Joseph. Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science. *Frontiers in Big Data*, 3, 2020. ISSN 2624-909X.

Acknowledgments

Pursuing a Ph.D. in AI within the last decade strikes me as being similar to navigating through the bustling traffic of Bombay. Both endeavors involve continuously evaluating evolving situations, adapting plans on the fly, and demanding a degree of composure in the face of uncontrollable circumstances. While the scientific community tends to celebrate the final results, these experiences have emphasized the significance of the journey itself. Or, as fellow fans of the epic fantasy genre might recognize, in the words of Brandon Sanderson, *Journey before Destination.* I would like to express my heartfelt gratitude to the many individuals who have played a crucial role in making my Ph.D. journey more fulfilling. When I arrived in the Netherlands in 2018, I found myself essentially starting anew, immigrating for the second time after having lived in the United States for seven years. Regardless of whether your name appears in these acknowledgments, if our paths intersected along the way, I extend my sincere thanks to you!

First and foremost, I would like to thank my parents. **Ma**, **Pa**, I am frankly amazed by how unwavering and unconditional your love and support is. My ability to pursue a career driven by curiosity is a privilege, one that has been sustained through sacrifices made by you. Thank you.

When the time came for me to choose between staying at CMU or embarking on a new journey at TUDelft to pursue my Ph.D., Alan Black had said to me, "You should go to Europe. They take the time to think about the right way of doing things, and you should experience it". **Hayley**, **Marco**, I have found that your approach to science embodies this ethos. I'm incredibly fortunate to have had the opportunity to collaborate with you, engage in enlightening discussions, and receive your support in what at times feels like a battle against the conventional ML/CV methods of tackling interdisciplinary research questions. **Marcel**, thank you for your wisdom and always being solution-oriented during our interactions. I appreciate the things you do behind the scenes to make the PRB section a productive and supportive working environment.

Meneer De Jong, thank you for humoring my crazy soldering requests without skepticism and helping with making my weird synchronization solutions a reality. **Bart**, I am delighted that you shared your interest in birds, insects, and Dutch wildlife with me. I enjoyed tracking the hawk that had taken shelter atop the Architecture tower with you through the surveillance camera you had hooked into. Together, you almost made me happy for facing IT issues, because that meant taking a break and walking over for a chat with you. **Saskia**, I find it amazing how you single-handle the work of at least five

people. PRB was lucky to have you. I'm glad you still stop by. Hope MMC is treating you well. **Marunka**, I apologize for being the first for you with several uniquely gnarly HR issues. I am glad that I can call you a friend despite these. You came in with an almost impossible task and I'm so happy to see you make the role your own. **Azza**, I appreciate your enthusiasm and initiative in making cluster usage a more civilized experience.

My Social BlaBla people, I truly appreciate the approach everyone in the lab has towards interdisciplinary research. I enjoyed all our meetings, especially the first hour. The others will never know the true answer to "But what do they discuss in such long meetings?" **Stephanie** and **Jose**, my fellow ConfLab partners, this was an absolute roller coaster of a ride, but I'm glad we did it together (and survived). **Ekin**, I miss you man. We need someone to compete with Tom's volume on the floor. **Laura**, I finally know the OSI layers. You were the first one to take me around the offices to introduce me to people when I started, a tradition I have tried to keep alive within the group. **Bernd**, my fellow plaid shirt done in companion, you are the DUDE, man. I'm glad that our intimate experience of sharing rooms with doorless glass bathrooms has resulted in my becoming your favorite aggressive extrovert. **Tiff** and **Chenxu**, it heartens me to see your endeavors in embracing the computational side coming from the social side. I wish you the best and am happy to be of whatever help I can. **Merle**, thank you for your joining and engaging with the study group. It was great to get to know you better at the HI retreat! **Navin** and **Bilal**, you suffered under my supervision. I'm sure it was worth it. You are welcome.

My PR BlaBla people, thank you for accepting me as one of your own. **David**, thank you for always being there to answer my questions in your unique animated style. **Jesse**, thank you for always being there to answer my questions in your unique zen style. **Bob**, I'm glad I got to interact with you before your retirement, thank you for sharing your wisdom with me. As you said, PhD students who are effective at time management have sufficient time to work on their own ideas. **Ojas**, take notes. **Merve**, my best co-teacher, thank you for the scented Turkish napkins. Looking forward to shaping GenMod with you. **Jing**, thank you for always sharing whatever useful information you come across, I appreciate it. **Gijs**, you recently started, welcome! Hope to get to know you better soon. I did find your Ph.D. thesis on one of the shelves and really enjoyed it (especially the aesthetics), before discovering it was yours. **Tom V.**, we started in neighboring offices but I'm glad that your voice makes it still feel the same despite the reorganization. Thank you for introducing me to the real Dutch culture and timeless Dutch hits like "Toeter Op M'n Waterscooter" and "Drank en Drugs". **Alex**, I've consistently appreciated the distinctive viewpoints you've contributed to discussions, whether it's drawing parallels between physics and machine learning, sharing training insights, and even highlighting how being able to take a punch to the face can benefit science. **Skander**, we should ride together more often. I would like to see the Canyon in action. **Mahdi**, looking forward to the PRB chess league. **Myrthe**, I

love how socially engaging your presentations are. Please never change that aspect. **Jim**, thanks for diving into XAI with me, our initial foray into the field served me well in my later projects. **Taylan**, thanks for handling the drinks for the borrel. I shall do my best to avoid them to reach your resting heart rate, you crazy beautiful fixie-riding machine. **Stephan**, I really appreciated our discussions while walking in the woods at the PR retreat. **Herr Karlsson**, we should really introduce the option of having Kaffe Karlsson during fika. Thanks for all the discussions on causality and productivity tools. **Ramin G.**, thanks for the violin lessons! **Jin**, **Yuko-san**, **Xiangwei-sama**, and **Mo**, thanks for all the tutorials on Mandarin and Japanese.

My CV BlaBla people, I enjoyed hanging out with you at conferences and poster sessions. **Jan**, I've gained a lot of wisdom from you over the years. I'm happy that we also got to know each other better on a personal level at ECCV. Hope to see you on the Rift sometime! **Nergis**, thank you for introducing me to some cool manga and anime. Please thank Çağrı for letting me slap your bass that one time. **Seyran**, I loved our discussions about interdisciplinarity. Looking forward to having more of them. **Silvia**, thank you for answering all my questions about Romania. I deeply appreciate that time when you went out of your way to get me the thread necklace from Amsterdam. **Xucong**, it's great to have someone with similar research interests. Looking forward to building the INSY Capture Lab system with you. **Attila**/**Robert-Jan**, I appreciate your clever use of holographic projections at ECCV to lure me into believing you are two different people. Nice try. I appreciated all our discussions on industrial versus academic research, and our shared passion for improved engineering practices in academia. **Ombretta** and **Aurora**, sincerely thank you for teaching me how to correctly pronounce Italian, and still putting up with me when I still do not. **Osman**, we should start playing football again so that it increases the odds of you finally scoring a left-footed curler into the top-left corner. I believe in you. **Yancong**, you are an inspiration in the gym, a terror on the football field, and the most wonderful person overall. I'm glad to call you a friend. **Ziqi**, it was wonderful sitting next to you at the start of our PhDs; I miss your Ziqi-isms. **Xiangwei-sama**, you are undoubtedly an *omoshiroi* person with many talents. Thank you for sharing them with us. **Hesam**, I was excited to learn you are into bikes too, we should ride together sometime. **Chengming**, I'm glad we now share a common appreciation of the emotional aspects of the data collection process. **Alejandro**, we haven't spoken much, looking forward to getting to know you better. **Sander**, I hope to one day match your barbequing skills. **Amogh**, I feel like we ought to have somehow interacted more during the course of the Ph.D., but it's always good to see you. Thanks for the Vicar Vision tour, restaurant recommendations, and multi-language LaTex tips! **Marian**, thanks for the introduction to Zettelkasten. It is great how you have something cool to share every time we speak, and I look forward to exploring 3D game engines with you. **Yunqiang**, **Xin**, you both have a quiet disposition,

but I enjoyed the conversations we had over the years. **Marcos**, fue un placer conocerte, espero que estés bien y nos veamos pronto.

My B-team amigos, **Yeshwanth** and **Arman**, thanks for diving head-first into my crazy cross-country biking ideas even though it was abundantly clear we were emphatically underprepared. Let's do it again. Also, thanks for agreeing to be my paranymphs. Yeshwanth, I find it rare to have a friend with whom I can talk about any topic for hours. You're my brotha from anotha motha. Arman, you are the OG doodool tala, macha. Thanks for helping with the Dutch translations of the summary. I look forward to many more discussions about electronics, equipment, and adventures.

My Bio BlaBla people, it was a pleasure to attend two full Biotalks over the four years, one even before Amelia arrived in Delft. **Ramin S.**, we started on the same day, and it has been an absolute joy discussing cinema, anime, video games, board games, and everything else with you, especially while eating the delicious food your various neighbors cooked for you. Looking forward to sharing more old Bollywood dance sequences with the music swapped. **Stavros**, we started off discussing statistical testing, but I'm glad our friendship developed since then to the point I could share a traditional Greek dance with your dad. It remains one of the highlights of my Ph.D. years. **Tamim**, thanks for your tough love (*cough* heckling *cough*) on the football pitch, I wouldn't have experienced the joy of scoring so many goals on a bad day without you. **Soufiane**, thanks for the discussions on VAEs and for introducing me to how you use them within the context of bioinformatics. **Christine**, it was always fun listening to your stories and gossip. **Christian**, I appreciated your humor, you made me want to know more people from Wageningen. **Tom M.**, I'm grateful someone else enjoys Calve as much as I do. Thank you for the board game sessions and all the help with the Dutch translation for the summary and printing tips. **Arlin**, you are so cool. I hope to attend one of your concerts someday. **Sally**, it was great playing football with you, for the few seconds I saw you before you sped past me (why were we always on opposite teams?) Thanks for enjoying my stories. **Aysun**, thanks for all the restaurant suggestions, authentic honey, and board game nights! **Meng**, I am impressed by all your cool gadgets for recording and projecting videos. **Mostafa**, seeing you deal with your injury on the football field, I can safely say you are one badass dude. Also, thanks for driving me during emergencies. **Colm**, I appreciate the educational enthusiasm with which you not only share your martial arts knowledge but also perform your rear naked choke holds. **Yasin**, I appreciate your social initiative and the effort you put into proposing fun activities for everyone. **Stephanie**, Amelia will share my acknowledgment since you like her more than me. **Chengyao**, thank you for being concerned for Amelia's safety when I was seated next to her at Meng's farewell. **Madelon**, you are hands down the most thoughtful designer of farewell gifts and an asset to the group. **Mo**, I'm sad that you do not appreciate the benefits of pytorch-lightning. **Akash**, thanks for introducing me to

the nuances of skydiving! **Paolo**, **Gabriel**, good luck with your VAEs! **Daniyal**, आओ कभी हवेली पे, पॉट लक किया जाय. **Sander**, I heard you biked to the Bio retreat. Good man. Please make this a thing for the rest of us. **Swier**, still waiting on the lifting tips, preferably delivered while you're rocking your rugged bearded look. **Jasper**, I'm grateful for your kindness in helping me make sense of the Dutch tax system and all their letters. **Gerard**, **Lieke**, I promise to visit the Dutch PhD office more often. **Kirti**, hope to see you more in the gossip room. **Paul**, you definitely have the second-best stories in the PRB. **Roy**, I promise I will never forget your name again. **Ahmed**, your transformation when an old Egyptian song comes on is magical. You are now my dancing idol. **Erik**, thanks for introducing me to the phrase *de vogel poept weer op dezelfde hoop.* I think it is a poignant description of academic grants. **Thomas**, thanks for hosting the cook-offs! I often argued with Yeshwanth about who won the cook-off, until we realized it had to be you given you got both dishes. **Joana**, I appreciate your efforts in organizing the tenure-tracker drinks and lunches. I'm sorry I haven't been able to join so far, but will rectify this soon! **Jana**, let's do some generative modeling together! **Jasmijn**, I'm inspired by your patience every time I invade your office.

Yuki-sama, **Jacopo**, **Fran**, **Rishabh**, **Nakisa**, **Yeshwanth**, **Aysun**, **Caro-chan**, **Willem**, **Lorena**, **Michael**, it was a pleasure serving on the EWI PhD Council with you all. **Ada**, **Sanne**, it heartens me to see that you genuinely care about the students. Thank you for fighting the good fight.

I would also like to extend my thanks to colleagues and friends from other departments. **Cynthia**, **Jorge**, I'm grateful for everything that you do to make the faculty a better place. I hope to continue working with you on these issues in my new role. **Odette**, you have been unfailingly kind and gracious to me during every interaction I have had with you over the years. Thank you. I hope to work with you soon in the near future. **Caro-chan**, the coolest thing about you is undoubtedly that you introduced me to Dávid. **Enrico**, I am yet to see someone pronounce Pei-Yu's name more accurately. **Pei-Yu**, thanks for putting up with us. **Masha**, I'm glad we got to know each other better at the HI retreat! **Sid**, eager to hear what's next for you! **Lukas**, one of these days I'll start joining the reading group more regularly. Thanks for running it so smoothly! **Chibuke**, you're a pure striker man, please pass some of your skills to Osman. **Taygun**, thanks for making me look like a better striker than I am. **Eric**, Quaresma ain't got nothing on your outside foot shots. **Oğuzhan**, you're the most silky passer on any team. **Alaeddin**, you are the most chill and fun teammate, thanks for all the laughs. **Amitabh**, many thanks for hosting me during my first days in the Netherlands, you made me look forward to being here. **Adnan**, the delightful evenings spent sipping chai and enjoying Coke Studio remain among my fondest memories from the first year of my Ph.D. Hope we can repeat them soon. **Ambareesh**, we should do a vintage-bikes-only ride sometime. Eindhoven isn't that far, is it? **Mythili**, I cherish our

friendship; it's unique to be able to discuss chess puzzles, other puzzles, Indian classical music, Escher, camping, specific differences between ducks and mallards, and hibiscuses and hollyhocks, oh and bikes (!), all with one friend.

To my extended family, whose initial excitement towards a potential career in medicine for me only deepened as they wholeheartedly embraced and supported my fascination with Computer Science, I am profoundly grateful for your enduring love and unwavering support. Thank you. બા, હું ડોક્ટર બની ગયો. Not the type you all were eager for, but still technically a Doctor. As we all know, technically correct is the best type of correct.

Lastly, **Amelia**, I want to express my deepest gratitude for being a steadfast companion throughout this journey. You've stood shoulder to shoulder with me in the trenches, cheered me on from the sidelines, and shared in both my setbacks and victories. You've been my invaluable sounding board, meticulous editor, LaTeX wrangler, figure optimizer, and 11th-hour savior under tight deadlines. But most importantly, you've invariably been there whenever I have needed it. This dissertation is as much yours as it is mine.

Chirag
Delfgauw, October 2023

Curriculum Vitæ

Chirag Anantha Raman

25-09-1988 Born in Mumbai, India.

Education

2018–2023 **Ph.D. Computer Science**
Delft University of Technology, The Netherlands
Promotors: Dr. H. Hung, Prof. dr. M. Loog, and Prof. dr. ir. M.J.T. Reinders

2011–2013 **Master of Entertainment Technology**
Carnegie Mellon University, USA

2006–2010 **Bachelor of Engineering, Information Technology**
University of Mumbai, India

Experience

2023–present **Delft University of Technology (TUDelft)**, Delft, The Netherlands
Assistant Professor, Department of Intelligent Systems, Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)

2021 **Microsoft Research**, Cambridge, UK
Ph.D. Research Intern, Presence AI

2016–2018 **Carnegie Mellon University - Language Technologies Institute**, Pittsburgh, USA
Senior Research Engineer (Jul '17 – Jul '18)
Research Engineer (Apr '16 – Jul '17)

2014–2016 **ProductionPro**, New York, USA
Lead iOS and UX Developer

2013–2014	**Disney Research**, Pittsburgh, USA Research Associate - Computer Vision
2012–2013	**Disney Parks, Experiences, and Products**, Orlando, USA New Technology Analyst - Next Generation Experience Project
2012	**Microsoft - User Experience and Creative Services**, Redmond, USA Developer, Project Wall# (Semester Project)
2011–2012	**Hungama Digital Media**, Mumbai, India Developer - iOS Games and Interactive Installations
2008–2011	**Indian Institute of Technology Bombay**, Mumbai, India Project Engineer - Project OSCAR (Jun '11 – Aug '11) Research Intern - Project OSCAR (Jul '08 – Jun '11)

Awards

2023	Google Initiated Grant (gift of 30,000 USD) for "A reliable framework for evaluating synthetic nonverbal behavior towards creating socially aware digital humans"
2023	First place (12 teams, 50 researchers), Emotion Physiology and Experience Collaboration (EPiC) Challenge, ACII 2023. Task: Inferring emotions (arousal, valence) from physiological signals
2022	Outstanding Reviewer, NeurIPS 2022 Datasets & Benchmarks Track
2015	(For ProductionPro) Audience's Choice Award, Demo Day, Made in New York Media Center
2012	To Innovation and Beyond, Walt Disney World New Media Group
2011	The Award of First Penguin, Entertainment Technology Center, Carnegie Mellon University
2011	K.C. Mahindra Scholarship for post-graduate studies
2011	Bharat Petroleum Corporation Scholarship for higher studies
2009, 2010	Sir Dorabji Tata Trust Scholarship for excellence in undergraduate studies

List of Publications

Journal

1. **C. Raman***, N. Raj Prabhu*, and H. Hung. Perceived Conversation Quality in Spontaneous Interactions. *IEEE Transactions on Affective Computing*, pp. 1-13, 2023. DOI: 10.1109/TAFFC.2023.3233950.
2. J. Vargas-Quiros, S. Tan, **C. Raman**, L. Cabrera-Quiros, and H. Hung. Covfee: an extensible web framework for continuous-time annotation of human behavior. *Understanding Social Behavior in Dyadic and Small Group Interactions, Proceedings of Machine Learning Research (PMLR)*, 173, pp. 265-293, 2022.

Conference

1. **C. Raman**, A. Nonnemaker, A. Villegas-Morcillo, H. Hung, and M. Loog. Why Did This Model Forecast This Future? Information-Theoretic Saliency for Counterfactual Explanations of Probabilistic Regression Models. *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, USA, 2023.
2. **C. Raman**, C. Hewitt, E. Wood, and T. Baltrušaitis. Mesh-Tension Driven Expression-Based Wrinkles for Synthetic Faces. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, Hawaii, USA, 2023, pp. 3515-3525.
3. **C. Raman***, J. Vargas-Quiros*, S. Tan*, A. Islam, E. Gedik, and H. Hung. ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild. *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, New Orleans, USA, 2022.
4. E. Wood, T. Baltrušaitis, C. Hewitt, M. Johnson, J. Shen, N. Milosavljević, D. Wilde, S. Garbin, **C. Raman**, T. Sharp, I. Stojiljković, T. Cashman, and J. Valentin. 3D Face Reconstruction with Dense Landmarks. *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, Tel Aviv, Israel, 2022, pp. 160-177. DOI: 10.1007/978-3-031-19778-9_10.
5. **C. Raman***, S. Tan*, and H. Hung. A Modular Approach for Synchronized Wireless Multimodal Multisensor Data Acquisition in Highly Dynamic Social Settings. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, Seattle, WA, USA, 2020, pp. 3586-3594. DOI: 10.1145/3394171.3413697
6. H. Hung, **C. Raman**, E. Gedik, S. Tan, and J. Vargas-Quiros. Multimodal Data Collection for Social Interaction Analysis In-the-Wild. *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*, Nice, France, 2019, pp. 2714-2715. DOI: 10.1145/3343031.3351320.

7. Y. Du, **C. Raman**, A. W. Black, L. P. Morency, and M. Eskenazi. Multimodal Polynomial Fusion for Detecting Driver Distraction. *Proceedings of Interspeech 2018*, Hyderabad, India, 2018, pp. 611-615. DOI: 10.21437/Interspeech.2018-2011.
8. S. Mehta, **C. Raman**, N. Ayer, and S. Sahasrabudhe. Auto-Grading for 3D Modeling Assignments in MOOCs. *IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, Mumbai, India, 2018, pp. 51-53. DOI: 10.1109/ICALT.2018.00012.
9. T. Y. Hu, **C. Raman**, S. M. Maza, L. Gui, T. Baltrusaitis, R. Frederking, L. P. Morency, A. W. Black, and M. Eskenazi. Integrating Verbal and Nonverbal Input into a Dynamic Response Spoken Dialogue System. *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'17)*, San Francisco, California, USA, 2017, pp. 5091-5092.

Workshop

🎓 1. **C. Raman**, H. Hung, and M. Loog. Social Processes: Self-Supervised Meta-Learning over Conversational Groups for Forecasting Nonverbal Social Cues. *European Conference on Computer Vision (ECCV) Workshop on Computer Vision for Metaverse (cv4metaverse)*, Tel Aviv, Israel, 2022.

2. M. Tsfasman*, A. Saravanan*, D. Viner*, D. Goslinga, S. de Wolf, **C. Raman**, C. M. Jonker, and C. Oertel. Towards a Real-Time Measure of the Perception of Anthropomorphism in Human-Robot Interaction. *Proceedings of the 2nd ACM Multimedia Workshop on Multimodal Conversational AI (MuCAI'21)*, Virtual Event, China, 2021, pp. 13-18. DOI: 10.1145/3475959.3485394.

🎓 3. N. Raj Prabhu, **C. Raman**, and H. Hung. Defining and Quantifying Conversation Quality in Spontaneous Interactions. *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion)*, Virtual Event, Netherlands, 2020, pp. 196-205. DOI: 10.1145/3395035.3425966.

🎓 4. **C. Raman** and H. Hung. Towards Automatic Estimation of Conversation Floors within F-formations. *8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, Cambridge, United Kingdom, 2019, pp. 175-181. DOI: 10.1109/ACIIW.2019.8925065.

Patent

Ⓡ 1. T. Baltrusaitis, **C. Raman**, C. Hewitt, and E. Wood. Face image generation with wrinkles.

* Equal contribution 🎓 Included in this Thesis Ⓡ Patent

"In order to make sustainable progress beyond leaderboards, AI/ML practitioners need to be more concerned about the details of the data they are working with, and we need more work on responsible data capturing approaches for "difficult" (e.g., hard-to-obtain, potentially sensitive) data. Your work is a valuable step in this direction."

— Ch. 2, Reviewer JhBB, NeurIPS Datasets & Benchmarks '22, Clear Accept

"The paper is mathematically very robust, with rigorous notation. Equations are always well presented and discussed. The idea of this approach is novel. Experimental validation is also rigorous."

— Ch. 5, Reviewer 1, CVPR '22, Weak Reject

"I like the way the authors grounded [the method] in the social science literature. However, I believe this made the paper hard to understand."

— Ch. 5, Reviewer 2, ECCV '22, Weak Reject

"This is a strong paper that is relevant to the community. The formulation seems quite elegant, and the results on both synthetic and natural experiments are strong."

— Ch. 6, Reviewer 82xZ, NeurIPS '23, Accept

"[The paper] is very well-written, well-structured and easy to read. It is opening a new way of looking at group dynamics."

— Ch. 7, Reviewer 1, ACIIW '19, Accept