

Sensing and Modeling Human Behaviors In Complex Conversational Scenes

Tan, S.

DOI

[10.4233/uuid:81a6854c-bc1d-4af5-816e-29a89ccc436d](https://doi.org/10.4233/uuid:81a6854c-bc1d-4af5-816e-29a89ccc436d)

Publication date

2023

Document Version

Final published version

Citation (APA)

Tan, S. (2023). *Sensing and Modeling Human Behaviors In Complex Conversational Scenes*. [Dissertation (TU Delft), Delft University of Technology]. <https://doi.org/10.4233/uuid:81a6854c-bc1d-4af5-816e-29a89ccc436d>

Important note

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

Copyright

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

Takedown policy

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

SENSING AND MODELING HUMAN BEHAVIORS IN COMPLEX CONVERSATIONAL SCENES

SENSING AND MODELING HUMAN BEHAVIORS IN COMPLEX CONVERSATIONAL SCENES

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board of Doctorates,
to be defended publicly on
Wednesday 3 May 2023 at 10:00 o'clock

by

Stephanie TAN

Master of Science in Computing Science, Imperial College London, UK
born in Pasadena, California, USA.

The dissertation has been approved by the doctoral committee:

Rector Magnificus,	chairperson
Dr. H. Hung,	Delft University of Technology, promotor
Prof. dr. ir. M. J. T. Reinders,	Delft University of Technology, promotor
Dr. D. M. J. Tax,	Delft University of Technology, copromotor

Independent members:

Dr. J. Ward,	Goldsmiths, University of London, UK
Dr. X. Alameda-Pineda,	The National Institute for Research in Digital Science and Technology (INRIA), France
Prof. dr. A. A. Salah,	Utrecht University, NL
Prof. dr. C. M. Jonker,	Delft University of Technology, NL
Prof. dr. P. S. Cesar,	Delft University of Technology, NL, reserve member

This research was partially funded by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.606.



Printed by: ProefschriftMaken || www.proefschriftmaken.nl

Cover by: ProefschriftMaken || www.proefschriftmaken.nl

Layout by: ProefschriftMaken || www.proefschriftmaken.nl

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

CONTENTS

Summary	ix
Samenvatting	xi
1 Introduction	1
1.1 Human Interactions	2
1.1.1 Social Signal Processing in Complex Conversational Scenes	2
1.1.2 Free-standing Conversation Groups in Complex Conversational Scenes	4
1.1.3 Social Signal Processing Approaches and Challenges.	5
1.1.4 Modeling Social Dynamics	7
1.2 Current Limitations	8
1.2.1 Sensing and Data Acquisition	8
1.2.2 Modeling Approaches	9
1.3 Addressing the Limitations - A Two-Pronged Approach	11
1.3.1 Sensing and Data Acquisition	11
1.3.2 Modeling Approaches	13
1.4 Contributions	14
2 Multimodal Joint Head Orientation Estimation in Interacting Groups via Proxemics and Interaction Dynamics	17
2.1 Introduction	18
2.2 Related Work.	21
2.2.1 Wearable Sensor Based Orientation Measurement and Estimation	21
2.2.2 Room Mounted Camera-Based Head Orientation Estimation	22
2.3 Methodology.	23
2.3.1 Approach	23
2.3.2 Implementation Details	25
2.3.3 Baseline Methods	26
2.4 Head orientations in complex social scenes: a case study.	28
2.4.1 Dataset.	28
2.4.2 Head Orientation Analysis	30
2.5 Results and discussion	31
2.5.1 Model Comparison.	32
2.5.2 Generalization to Unseen Data	33
2.5.3 Contribution of Different Modalities	34
2.5.4 Regression vs. Classification	34
2.5.5 Using Body Orientations	35
2.5.6 Speech Dynamics vs. Head Orientation Estimation.	36

2.6	Discussion and Future Work	37
2.6.1	Source of Inputs	37
2.6.2	Importance of Prior Knowledge	37
2.6.3	Multimodal Features and Fusion Methods	38
2.6.4	Group-Size-Agnostic vs. Group-Size-Specific Models	38
2.7	Conclusion	38
3	Conversation Group Detection With Spatio-Temporal Context	39
3.1	Introduction	40
3.2	Related Works	42
3.3	Approach	44
3.3.1	Affinity Prediction	45
3.3.2	Dominant Set Clustering	47
3.4	Experimental Setup	47
3.4.1	Baseline Methods	47
3.4.2	Datasets	48
3.4.3	Evaluation Metrics	48
3.4.4	Implementation Details	49
3.5	Results and Discussion	50
3.5.1	Overview	50
3.5.2	Analysis of Affinity Values	51
3.5.3	Affinity Scores in Dominant Set Clustering	51
3.5.4	Performance With Respect To Scene Dynamics	52
3.5.5	Conversation Group Forecasting	53
3.6	Conclusion and Future Works	54
4	Head and Body Orientation Estimation with Sparse Weak Labels in Free Standing Conversational Settings	55
4.1	Introduction	56
4.2	Related Work	57
4.2.1	Human Pose Estimation	57
4.2.2	Head and Body Orientation Estimation: RGB Data	58
4.2.3	Head and Body Orientation Estimation: Depth and Wearable Sensors	59
4.3	Overview of the Approach	59
4.4	Proposed Model	60
4.4.1	Rank Minimization	62
4.4.2	Temporal Smoothing	62
4.4.3	Regularization by Weak Labels	64
4.4.4	Head and Body Coupling	64
4.4.5	Optimization problem	64
4.5	Experiments	65
4.5.1	SALSA Dataset Analysis	65
4.5.2	Experimental Setup	67

4.6	Model Analysis.	68
4.6.1	Results	68
4.6.2	Kernel Choice	69
4.6.3	Regularization by Weak Labels	70
4.6.4	Contribution of Head-Body Coupling	71
4.7	Discussion and Conclusion.	72
5	A Modular Approach for Synchronized Wireless Multimodal Multisensor Data Acquisition in Highly Dynamic Social Settings	75
5.1	Introduction	76
5.2	Related Work.	77
5.3	Our Approach	80
5.3.1	NTP as A Reference Signal	80
5.3.2	Real-world Implementation	82
5.3.3	Latency Measures in Social Literature	84
5.4	Experiments	85
5.4.1	Timecode Latency between NTP-LTC Converter and Camera Network Master	85
5.4.2	Evaluating Crossmodal Synchronization	88
5.5	Cost versus Latency Considerations	89
5.6	Conclusion.	90
6	ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild	91
6.1	Introduction	92
6.2	Related Work.	94
6.3	Data Acquisition	95
6.4	Data Annotation	97
6.5	Dataset Statistics.	99
6.6	Research Tasks.	100
6.6.1	Person and Keypoints Detection	100
6.6.2	Speaking Status Detection	101
6.6.3	F-formation Detection	102
6.7	Conclusion and Discussion.	102
7	Discussions and Conclusion	105
7.1	Accurate Perception of Human Interactions In-the-Wild	107
7.2	The Scale of Modeling Human Behaviors.	108
7.3	Concerns for Wearable Sensors.	110
7.4	Reproducible Data Collection	110
7.5	Modeling Human Interactions: A Different Viewpoint	111
	Bibliography	113
	Appendix	141
	Acknowledgments	165
	Curriculum Vitæ	167

List of Publications**169**

SUMMARY

Understanding human behavior has been an intriguing topic studied by many disciplines, including social science, neuroscience, etc. Humans exhibit social behaviors, through for example, interacting, conversing, empathizing with each other. Systematically and scientifically studying these behaviors often requires granular observations and measurements. With increasing digital sensor and computer sensing and processing capability, accurately measuring and recording large amount of real-life human social behavior has become possible. Computational methods, such as machine learning, can be developed to analyze these data in unprecedented ways by detecting and learning patterns in the signals. However, even with the available data and advanced machine learning methods, understanding human social behavior is still challenging, as it is contextual and could result in variations.

This thesis focuses on analyzing human behaviors in complex conversational scenes. It proposes novel computational methods that incorporate the context, which is the conversation group and the interaction scene. Prominent behavioral cues in social interaction include head and body orientations, as they are proxy indicators for visual attention and conversation group membership. This thesis first covers methods for head and body orientation estimation (under data-scarce and data-rich settings), and conversation group detection. These methods have an emphasis on learning from multimodal data and context modeling, and their efficacy is shown empirically. Then, the thesis addresses an open challenge in acquiring human social data in real-life by proposing an accurate and scalable method for data synchronization. Lastly, this thesis introduces a new dataset collected by the aforementioned synchronization method, capturing real-life interaction in a conference settings. Therein, results of tasks such as keypoint detection, action recognition, and conversation group detection are reported, which also motivate future research in this area. Combining these contributions in both computational method development and data collection, this thesis takes a step forward in understanding human behaviors in conversation scenes.

SAMENVATTING

Het begrijpen van menselijk gedrag is een boeiend onderwerp en wordt dan ook door vele disciplines zoals bijvoorbeeld sociale wetenschappen of neurowetenschappen bestudeerd. Mensen vertonen sociaal gedrag door bijvoorbeeld met elkaar te communiceren, te converseren of zich in te leven. Het systematisch en het wetenschappelijk bestuderen van deze gedragingen vereist vaak zeer gedetailleerde waarnemingen en metingen. Met de toenemende mogelijkheden van digitale sensoren en computerondersteunde detectie en verwerking is het nauwkeurig meten en het vastleggen van grote hoeveelheden menselijk sociaal gedrag in reallife mogelijk geworden.

Computationale methoden, zoals bijvoorbeeld machinaal leren, kunnen worden ontwikkeld om deze data op ongekende manieren te analyseren door patronen in de signalen op te sporen en hier van te leren. Maar zelfs met alle beschikbare data en geavanceerde methoden voor machinaal leren blijft het begrijpen van menselijk sociaal gedrag een echte uitdaging omdat het contextgebonden is en in kleine variaties kan resulteren.

Deze thesis richt zich op het analyseren van menselijk gedrag in complexe gesprekssituaties. Het stelt nieuwe computationele methoden voor die rekening houden met de context, te weten de gespreksgroep en de interactieve omgeving. Prominente gedragsaanwijzingen in de sociale interactie zijn de positie van het hoofd en de lichaamsoriëntatie omdat dit proxy-indicatoren zijn voor de visuele aandacht en het deel uitmaken van de gespreksgroep.

Deze thesis behandelt allereerst de methoden voor het inschatten van de hoofd- en lichaamsoriëntatie (op basis van dataschaarse en datarijke settings) en detectie en verloop van de gespreksgroep. Deze methoden leggen de nadruk op de manier waarop multimodale data het leren en de contextmodellering kunnen ondersteunen. De doeltreffendheid hiervan zal in deze thesis empirisch worden aangetoond. Vervolgens behandelt deze thesis de enorme uitdaging van het verwerven van menselijke sociale data in reallife door een nauwkeurige en schaalbare methode voor datasynchronisatie voor te stellen.

Tenslotte introduceert deze thesis een nieuwe dataset, verzameld door de bovengenoemde synchronisatiemethode, die de reallife interactie tijdens een overleg vastlegt. Daarin worden de resultaten gerapporteerd van taken zoals de detectie van kernpunten, de herkenning van acties en detectie van gedrag tijdens het bijwonen van gespreksgroepen, die ook meteen de motivatie vormen voor toekomstig onderzoek op dit gebied. Door deze bijdragen voor zowel de ontwikkeling van computationele methoden als het verzamelen van data te combineren, kan deze thesis een grote stap voorwaarts betekenen voor het beter begrijpen van menselijk gedrag in gesprekssituaties.

1

INTRODUCTION

1.1 HUMAN INTERACTIONS

Human beings are inherently social [1]. A large number of behaviors occur in the context of commonplace social interactions, including greeting, conversing, negotiating, turn-taking, and group-forming activities [2]. Even in this digital age of the world, face-to-face interactions remain relevant as a powerful way of communication that carries a rich collection of visual, vocal, and verbal information, where humans see, hear, and interact with each other. There is a rich history in studying human interactions, particularly through a social science lens (e.g., [3]). However, to build more socially-aware automatic systems for application such as conversation agents and human-robot collaboration, we could benefit from the sensing capabilities and modeling approaches to date. The automated detection and perception of human behaviors in the real world under naturalistic settings, which provide quantitative and more objective measurements of how humans behave and interact, is a feasible approach towards obtaining more insights. This moves away from the traditional approach (such as ethnography) rooted in sociology and psychology disciplines [4]. Hence, this thesis focuses on machine perception of human behavior (specifically in small group interactions) and makes advances at the intersection of computer science, engineering, and social science in this setting.

How humans interact with each other and how their behaviors manifest could drastically vary in different social scenarios [5]. For background, we first explain a framework to characterize different social settings. Dimensions to consider include (i) unfocused vs. focused (commonly or jointly) encounters, (ii) group size, (iii) static vs. dynamic organization, (iv) scripted/staged vs. in-the-wild. The nature of each setting carries corresponding implicit assumptions and social norms. Figure 1.1 extends the framework summarized by Setti et al. [6] by dividing different types of social scenes with respect to the presence of groups. We illustrate some examples of representative settings along these dimensions. Particularly, in taxonomizing groups, we use the definition of "group", a type of social entity that is often called the *small group*. In small groups, feelings of "groupness" occur in settings in which participants temporarily involve in shared and coordinated activity [7].

Goffman [8] conceptualized the difference between focused and unfocused interactions, where focused ones involve people who show involvement and pay visual and cognitive attention to each other. The distinction of commonly focused and jointly focused encounters is further explained by Kendon [9], where the latter describes settings that are more open and conducive to interpersonal conversation without a single common focus of attention (e.g., movie theatre). This thesis is situated at the last row of social settings in Figure 1.1, usually found in cocktail parties, networking events, etc.

1.1.1 SOCIAL SIGNAL PROCESSING IN COMPLEX CONVERSATIONAL SCENES

The formal study of nonverbal communication and social signals originated from social science. However, more recently, Vinciarelli et al. [10] proposed studying social signals through automated methods and coined the term, social signal processing (SSP). Through multimodal sensing and analysis of human behaviors, we could endow machines with the ability to comprehend and express social signals to achieve social intelligence. Social signals include various non-verbal behavioral cues which could be defined by (i) physical appearance, (ii) gesture and posture, (iii) face and eye behavior, (iv) vocal behavior, and lastly (v) space and environment (i.e., the way people share and organize the space they







Examples	Setting	Focus of attention	Scene change	Scene description
	Waiting room	unfocused	Static	No group
	Queue	unfocused	Dynamic	No group
	Movie theater	Commonly focused	Static	One large group
	Parade	Commonly focused	Dynamic	One large group
	Meeting	Jointly focused	Static	One group of variable size
	Networking event	Jointly focused	Dynamic	Multiple small groups of variable size

Figure 1.1: Examples of different social scenes (extended from [6]). For completeness, different examples of interaction scenes are shown. This thesis focuses on the last row, and more specifically, for conversational interaction.

have at disposition [10], which could affect their relationship with each other [11]. Upon understanding these behavioral cues, high level and downstream social behaviors such as turn-taking, dominance, empathy, etc. can be studied, especially with the aid of automated perception and analysis through modern sensing and machine learning methods.

Out of these behavioral cues, face and eye behaviors express social signals with highest effectiveness as they are direct and naturally preminent means of communicating [10]. Important social cues such as focus of attention in social settings could be inferred from eye gaze. Additionally, space and environment also directly shapes how people arrange themselves and interact. Intuitively, if the space is small (and more crowded) or restricted by furniture, the interpersonal dynamics would be affected. This can be best described by Edward Hall's theory on proxemics, which could be characterized by location, interpersonal distance, and orientation [11], as well as Kendon's theory on F-formation shaping [12].

Apart from these proxemics characteristics, **complex conversational scenes** are unstructured interaction scenes for which organization is not only spatially constrained but also socially driven. These scenes exhibit behavioral and conversation dynamics involving multiple members [13] and multiple groups that are physically co-present. During mingling (i.e. an example of complex conversational scene), the close interpersonal distance enables interaction within personal and close social space [11] which creates unique social dynamics that is not commonly observed in other social situations.

To capture behavioral cues in these complex conversational scenes, we rely on mul-

timodal sensors. We aim to measure and record human behaviors unintrusively while preserving ecological validity for in-the-wild settings. It becomes increasingly difficult, if not impossible, to capture behaviors such as eye gaze and facial expression for inference of focus of attention, compared to structure settings such as meetings.

Because of its importance to downstream and socially relevant studies (i.e., conversation quality/engagement estimation [14], modeling of interaction processes [15], etc.), the task of estimating head and body orientations is one of the main focuses of this thesis (Chapter 2 and 3). Head orientation is an important cue for social attention, particularly in the absence of eye gaze measurements [16]. Body orientation is an indicator for conversation group membership [12]. The incorporation of temporal and conversation dynamics is essential when modeling these behavioral cues because social interactions often involve multiple people and are constantly evolving. Subsequently, the resulting head and body orientations could serve as informative cues for the task of pairwise interactant affinity estimation (Chapter 4) which is the building block for extracting interaction groups) [17]. Once interaction groups are automatically detected, we could track conversation group evolution (breaking, forming, reforming, etc) and ultimately obtain more understanding of social scenes. Quantities such as head and body orientations, as well as pairwise affinity, constitute the main ingredients of social signals in mingling scenarios, that could enable further analysis, understanding, and synthesis of social behaviors.

1.1.2 FREE-STANDING CONVERSATION GROUPS IN COMPLEX CONVERSATIONAL SCENES

Most existing research on social behavior analysis focused on pre-arranged, staged, or scripted settings (specifically for commonly focused encounters) such as meetings with a small number of people (typically no more than six) [18–21]. The fixed nature of such settings and the low number of people do not reflect the complexity of the types of interactions that arise during mingling events such as networking events, which are more spontaneous and ad-hoc. The number of participants in a conversation group is not restricted and could change dynamically. To model social behaviors of individuals and groups more systematically, **free-standing conversation group (FCG)** was introduced and defined as a jointly focused interaction group during a social occasion, such as a party, event, or next to the coffee machine [22]. An FCG can best be formalized and described by Kendon’s definition of *facing-formation* or *F-formation*, which is socio-spatial formation in which participants of the same formation have and maintain a convex space (o-space) to which all participants have direct and equal access [23] (See Figure 1.2). F-formation is one of the most fundamental concepts of interaction in an interaction scene. While it is possible to infer interacting partners purely through social dynamics (e.g., body motion) [24], the identification of the F-formation is primarily driven by the participants’ locations and orientations (i.e., proxemics).

Studies that model human behaviors in FCGs rely on the unique properties of the social interaction scene that cannot be ignored or exactly approximated with another type of social scenario (e.g., lab-based) since insights and conclusions may not be transferable. Laidlaw et al. corroborates this point of view through empirical evidence that suggests the mere opportunity and presence of social interactions in the surrounding could alter where people look and how they behave [25]. In this thesis, we argue that it is important

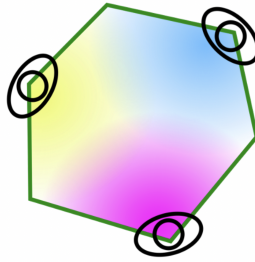


Figure 1.2: Visualization of an F-formation. Each subject has individual transactional segment. All of the subjects maintain a shared interaction in the middle (o-space).

to develop specific models for these scenarios rather than retrofitting pre-trained models or use data from other contexts for our use case to solve the automated inference problem.

1.1.3 SOCIAL SIGNAL PROCESSING APPROACHES AND CHALLENGES

To study behavioral cues in these conversation groups, we adopt the general approach common in SSP. A high level breakdown of tasks involves (1) data collection and preprocessing (2) extraction of informative behavioral cues (e.g., head and body orientations, speaking status), and (3) interpretation from behavioral cues, social signals, and/or social concepts by incorporating context-awareness, leading to social behavior understanding. This thesis contains work in data collection, and the automatic extraction of behavioral cues such as orientations and detection of social concepts such as F-formations. This thesis focuses on the boxed region of Figure 1.3.

Challenges related to the data collection include privacy and ethical concerns, ecological validity, and measurement fidelity. The design choice is often a trade-off among these considerations. Data collection involving human subjects studies needs to be carefully planned and executed to ensure privacy by addressing ethical concerns. Approvals from ethical board of institutions, in which details regarding informed consent, anonymization, data sharing, etc., must be documented beforehand. For example, collecting biometric information (e.g., facial images) and high frequency audio for transcription of verbal content is more sensitive compared to the recording of body movement. Therefore, sensor choices used in this type of data collection and their placement require strategization for the trade-off between unintrusiveness and fidelity, while preserving ecological validity. More specifically, we chose to use elevated side-view or top-down view, as opposed to frontal or egocentric view, which is more intrusive. Wearable sensors can be specifically designed and placed to reduce participants noticing their presence. As opposed to wearable sensors like full body suits, smart garment, etc., we chose to use single-worn wearable sensors (e.g., around the neck) to capture body movement without the participants noticing. With less intrusive sensors (both the camera view and placement of wearable sensors), participants acclimatize to the scenarios and quickly forget that they are being recorded, thereby maximizing ecological validity [26].

To extract behavioral cues from related upstream tasks from the data, relevant tasks for automated methods include person detection, person re-identification, speaker diarization, etc. These are active research topics in computer vision and speech communities, but

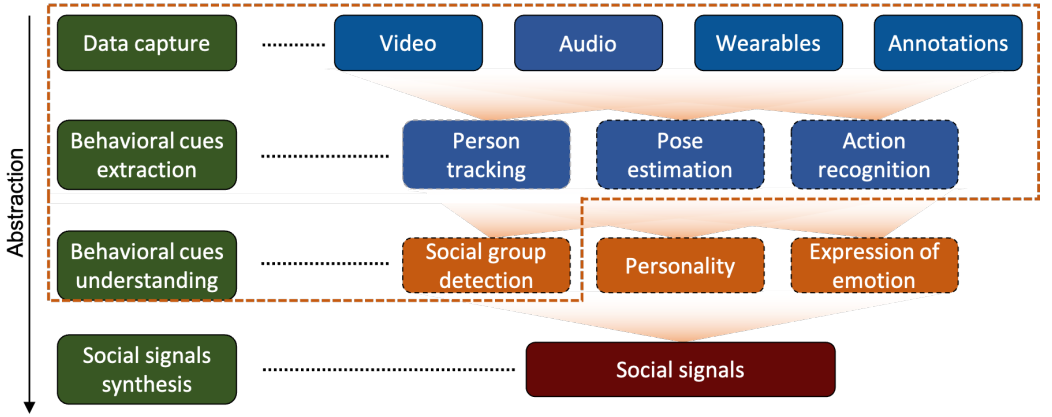


Figure 1.3: Social signal processing approach. This thesis includes works related to topics in dashed box.

remain challenging in our setting. Since the top-down viewpoint in the data is relatively uncommon compared to elevated side-views, and audio signals capture cocktail party noise in social interactions, off-the-shelf methods do not work directly and do not give results well enough for SSP researchers to employ downstream tasks. Hence, we often resort to using manually labelled data instead of automatically acquired ones to remove some upstream uncertainties and errors that could propagate to downstream tasks.

As motivated in Section 1.1.1, this thesis focuses on the modeling of posture, more specifically, head and body orientations, with context-awareness encoding space and the surroundings, not limited to location but also the context in social dynamics (i.e., conversational dynamics which arises from the multiple levels of coordination during conversing such as converging on speaking patterns and speech rate, speech rhythm and postural synchrony, etc. [27]). The methods proposed in this thesis incorporates multimodal data such as body motion measurements and speaking status in complex conversational scenes for modeling human head and body orientations, which constitute an extension of the previous works [28, 29], by accounting for the inherent social dynamics in-the-wild through the more flexible deep-learning based methods.

The steps leading up to social behavior understanding are necessary prerequisites for tasks such as modeling social relationships (e.g., role recognition, conversation group membership) and social attitudes (e.g., emotion, expression recognition). These high-level social phenomena are complex and rely on context (e.g., environment) and the nature of social interactions (e.g., formal vs. informal). Role recognition such as speaker/non-speaker identification could be achieved using video, audio or multimodal approaches [24], where speaking behavior such as turn-taking, interjections, overlaps, etc. serve as important cues. Modeling other social relationships like leader-follower is less formalized, and is limited to social science theories in group dynamics and processes [30]. Social attitudes such as dominance [31, 32], deception [33, 34], etc. are group-level behaviors. Past works in dominance estimation utilized speech activities, along with body movement and gaze. Additionally, there has been works in negotiation, rapport [35], agreement, among other socially relevant tasks. While there are many possibilities of social concepts that can be

modeled, this thesis focuses on studying conversation group membership through pairwise relationship (Chapter 4).

The expanded walk-through of the workflow of social signal processing shows that processing and understanding social signals involves multiple steps. It is difficult to combine them into a single end-to-end procedure, especially when different research fields are focusing on subsets of the tasks. Some works have expanded into modeling social relationships (e.g., looking at each other [36]) using behavioral cues such as gaze. However, when in complex and crowded conversational scenes, extracting behavior cues is still challenging since phenomena such as social relationship and attitudes could be related to context [37], and data including annotations could be especially subjective, when the concepts are abstract, involving the individual's cognitive perception of the situation (e.g., engagement), and not directly observable.

From extracting behavioral cues and modeling social concepts, we can then begin to understand the underlying social signals for building truly socially-aware perceptive systems. Social signals associated with seemingly intuitive concepts such as conversations are not yet fully understood, and ones for advanced concepts such as empathy, flirtation, etc. remain to be open research topics. However, we argue that the data and methods/tools resulting from the workflow, as well as the insights gained are crucial in achieving social awareness of automatic approaches.

1.1.4 MODELING SOCIAL DYNAMICS

Social dynamics (i.e., the underlying behavioral processes) is a complex and challenging phenomenon to model. We can begin to unpack some aspects of the social dynamics by associating speech and prosodic activities with conversation dynamics and body movement activities. Speaking status of individuals indicates turn-taking which reflects the conversation dynamics. Past works in conversation analysis model the temporal changes with dynamic Bayesian networks and other Markov models [38, 39]. Additionally, it has been found that interactions contain interpersonal coordination (i.e., coordinated bodily movements of co-actors in time) [40]. It is necessary to take into account the social dynamics when modeling human behaviors such as head and body orientations during social interactions.

Concretely, we argue and show that we could capture some aspects of the social dynamics through modeling temporal dynamics using multimodal signals. With the advancement of using sensors, we are able to measure human behaviors continuously and granularly. Signals from video, wearable sensors, or even labels through time have inherent temporal dynamics. Irregular variations in time series, which are nonrandom sources of variations, could be particularly interesting in the context of social dynamics which encapsulates a system of behavioral and psychological processes occurring in or between social groups. In face-to-face interactions, these variations could be manifested as, for example, changes in head/body orientations during a period of irregular activities (sudden movements resulting from changes in the social scene) relative to trends on a larger time scale. More generally, the individual behavioral time signals, with joint consideration of signals from others in the social groups, make the very fabric of the underlying social dynamics which is ultimately what automatic methods need to capture. Importantly, the interpretation of observed or predicted social behavioral cues need to be situated in the temporal dimension and its

dynamics [10].

As opposed to making assumptions (e.g., wide-sense stationary processes) for autoregressive approaches such as Autoregressive Moving Average methods (ARMA, ARIMA, etc.), we opted for developing deep learning approaches for time series modeling are not based on statistical assumptions and could serve as arbitrary approximators. While traditional approaches are still valid, using neural networks for time series modeling, especially in the multivariate case with multidimensional signals from multiple modalities in modeling human behaviors, offers flexibility and simplifies the modeling process for practitioners.

Similar to previous works [21, 28], we motivate the use of multimodal data (audio for speech activities, and/or accelerometers/inertial measurement units (IMUs) for body movements) to complement information that are not available in unimodal (image/video) data, which has been the standard approach in the past. This thesis investigates the feasibility of deep learning sequence-based methods using this combination of sensors to model social dynamics that occur specifically in settings like FCGs.

1.2 CURRENT LIMITATIONS

1.2.1 SENSING AND DATA ACQUISITION

Sensing human behaviors has been a long standing challenge for the social signal processing community. The technical and logistical difficulty is coupled with the need for maximizing ecological validity (i.e., ensuring the naturalness in the data acquired), as well as for accommodating data streams of multiple modalities. Many datasets contain recordings of human pedestrian behavior for person detection and tracking. Unfortunately, they are insufficient for modeling head and body orientations in social situations since the setting is not focused [9] and the location of subjects in crowded scenes is largely static, without a motion cue. Therefore, the need for developing new datasets was evident.

In most cases, video and audio sensors are affixed to the ceiling (e.g., [41, 42], table (e.g., [43, 44]), etc. When these sensors are in direct line of sight of the human participants, they introduce some bias towards how humans act under observation. Fortunately, subjects tend to acclimatize to the setting if they get sufficiently comfortable with the recording setup. For visual data, existing datasets for social interactions and human behaviors utilize elevated side-views [22, 45, 46] and top-down views [41, 42] for avoiding concerns of ecological validity, while egocentric ones are more intrusive (e.g., [47]). From the elevated side-view, automated methods such as person detection and tracking have been employed to further derive visual features from head and body bounding boxes. However, for elevated side-view cameras mounted in the scene (though unintrusive), there could be considerable occlusions which cause missing visual information. This could be alleviated by using a multi-view elevated-side view setup but requires choosing the best viewpoint manually [45]. Additionally, to compensate for the loss of visual information (e.g. occlusions), infrastructure such as the Panoptic dome [48] was built and equipped with hundreds of cameras. 3D human poses could be extracted to directly infer 2D head and body orientations. However, such infrastructure is impossible (cost and logistic efforts) to replicate at real-life events, compared to a smaller number of cameras/sensors capturing the whole interaction scene.

For audio data, microphone arrays placed on the meeting table have been used to record

meeting interactions. Personally worn microphones have been incorporated in wearable smart ID badges along with other sensors. While microphones used in more structured settings are also used to capture audio information (typically at high sampling rates) for tasks such as speaker diarization, wearable microphones in smart ID badges have been adjusted to operate on low sampling frequency for privacy preservation. The types of data that could be derived include speaker identification, prosody, turn-taking, etc., but not necessarily verbal transcription.

Body motion is traditionally captured using full body motion capture, and typically involves a fully calibrated multi-camera setup and wearing a sensor suit. This setup is expensive and in practice difficult to fit in with data streams of multiple modalities, especially from the synchronization perspective in terms of signal alignment. New technology and sensors focus on sensor minimization and do not require a physical connection from the suit to the data capturing system. Figure 1.4 shows examples of representative complex conversation scene and the set of wearable sensors that have been used for data collection. In the wearable smart ID badge form-factor (see Figure 1.4(d)), body motion is captured using an inertial measurement unit (IMU), which contains accelerometer, gyroscope, and magnetometer. IMU units are commonly found and inexpensive. Mobile phones and smart bracelets/watches also contain an IMU sensor. For scenes shown in Figure 1.4, wearable smart ID badge worn around the neck can be deployed in a conference setting and other similar in-the-wild settings. They incorporate different sensors able capture speech and torso activities, and serve as an alternative sensing method than mobile phones.

While these design choices for each individual modality are justified, the integration of multimodal streams for data acquisition in-the-wild is a non-trivial task. Because of the standing technical challenges and monumental logistical effort in data collection, there is only a handful of datasets (with small sample size) for complex conversational scenes. This hinders the development of automated methods.

1.2.2 MODELING APPROACHES

HEAD AND BODY ORIENTATION ESTIMATION

Head and body orientation estimations have been extensively studied in the past in the computer vision community. Representative works [28, 49, 50] focus on using elevated side-views in an 8-class classification setting. More recently, there has been advances in 2D pose estimation methods that detect human skeletal keypoints. Orientations cannot be directly extracted from these detected 2D keypoints without an indication of directional vector. Works related to the Panoptic studio [18, 51] have relied on 3D body motion and face motion outputs to compute the body orientation and face orientation by finding the 3D normal vector direction of the torso and face. However, without overlapping views of the scenes using RGB cameras, 3D information is difficult to reconstruct. Importantly, these existing methods on head and body orientations mostly rely on visual facial information or motion cues and they are not directly applicable to the top-down view (as shown in the keypoint estimation task Chapter 6).

Some previous methods capture temporal dynamics in modeling conversation patterns through dynamic Bayesian network models [52] and Markov models up to an order [39]. The complexity of these approaches increase dramatically when longer temporal context is considered. Other models that take the temporal patterns into context such as ones

proposed by [45] and [53] do not consider the interaction dynamics, such as speaking activity or fine grained body movement explicitly. This points to the advantage of novel methods, as ones included in thesis, to model for the temporal dependency in the inputs.

Multimodal modeling of head and body orientations have been studied in [28, 54], using a combination of video, microphones, and wearable sensors like smart ID badges. Labels of head and body orientations could be inferred from these sensors, though they could be less reliable than ones obtained from videos, especially ones obtained through multiple views (see Chapter 3). Aside from ensuring the temporal smoothness in the approaches, there has not been any explicit modeling of the temporal context of the task for head and body orientation estimation. Social context also has not been explicitly accounted for, except in [29, 50] primarily using the visual modality. However, we argue that social context in which we express ourselves can be better captured with the additional modalities such as speech (via speaking status) and body movement (via acceleration), to account for conversation dynamics and movement synchrony patterns, respectively.

CONVERSATION GROUP DETECTION

Social relationships and attitudes arise generally in the presence of interactions. Here we focus on automatic detection of conversation groups. Knowing who is talking to whom, and detecting who is interacting with whom is a critical step from modeling individual behaviors to group behaviors, in order to understand more high level social phenomena. There is a feedback relationship between behavioral cues and interaction dynamics: cues like head and body orientations are social signals that decide the interaction dynamics, and the generated dynamic also affects how people display subsequent behavioral cues (i.e. how people responded to the process of interaction)[55–58].

Formulating this concept into one that can be modeled automatically has been a challenging. Past works such as [22] model the o-space explicitly. Other works such as [17] constructed affinity matrix of the scene and developed a clustering approach based on Dominant Set for conversation group extraction. More recent deep learning approaches (e.g., [59, 60]) have made significant improvements in performance in typical datasets for this task (e.g., [45, 46]). However, these approaches do not model the temporal context, and hence miss capturing part of the underlying social dynamics. The main types of dynamics in human face-to-face social interactions in networking scenes are the changing proxemics (i.e., location and orientations), conversation dynamics, and body coordination dynamics. The evolving landscape of the interaction scenes is determined by how people move around each other to a large extent. In this thesis, we tackle this task by accounting for the spatio-temporal context in detecting conversation groups (Chapter 4).

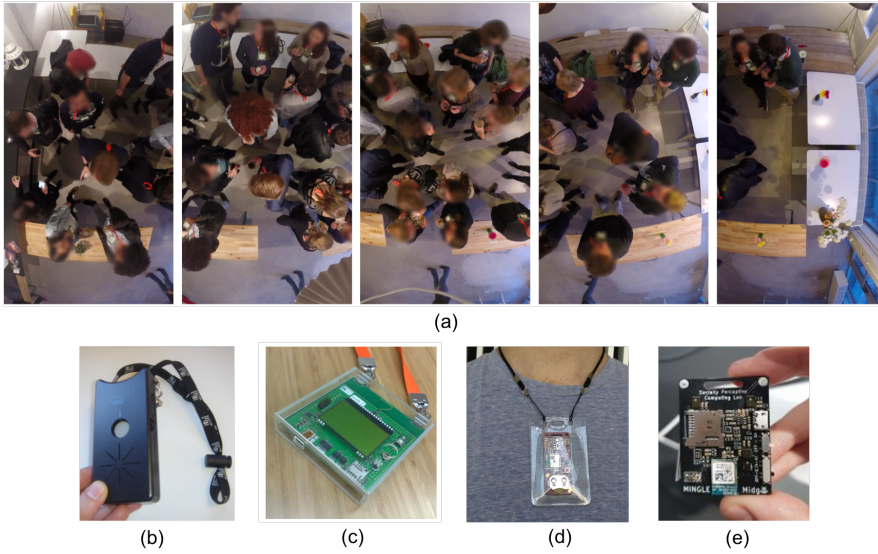


Figure 1.4: Typical setup of data collection experiments of complex conversation scenes with different sensor options. (a) shows a typical overhead top-down view of the interaction scene [41]. (b)-(e) exemplifies wearable sensors that resemble a conference badge form-factor. (b) shows Sociometric badge [61]. (c) is the chalcedony badge [62]. (d) is the rhythm badge [63]. (e) is the midge badge.

1.3 ADDRESSING THE LIMITATIONS - A TWO-PRONGED APPROACH

This thesis aims to address these aforementioned challenges, which are separated in two broader categories: (1) related to data acquisition (i.e., data collection and annotations), and (2) related to modeling.

In modeling social behavioral cues, it is crucial to consider the availability and quality of data, as behaviors are context-sensitive. High quality data (high fidelity, fine-granularity) with reliable labels are scarce, difficult and expensive to acquire. While researchers are urged to evaluate the amount and variety of data required, as well as requirements of generalization, this thesis proposes an interpretation and arrangement of the works applicable under the following scenarios: (i) when there is available data (Chapter 2,3), (ii) when there is low data availability (Chapter 4), and (iii) when there is no data (Chapter 5,6). In this thesis, modeling-related works focusing on head and body orientation estimation and group detection are situated in the first two scenarios in light of general multimodal data scarcity and working with existing datasets. Sensing-related and data collection works are situated in the third scenario, with the goal of enabling future research on social behavior understanding and synthesis.

1.3.1 SENSING AND DATA ACQUISITION

Data capturing, related to sensor perception, synchronization, and annotation acquisition have been a challenge in the social signal processing field. This is illustrated by the scarcity

of human social behavior dataset captured in in-the-wild settings (unscripted/unstaged), as data collection experiments require extensive logistical coordination and annotations are costly and labor-intensive [64, 65]. As motivated in Section 1.3.1 (data sources and fidelity), capturing human behaviors involve using multimodal sensors. Previously in more structured settings like meetings with fewer participants, the acquisition method has relied on physical connections to a computer node (e.g., a multi-channel audio interface) where all the sensor data is recorded [66, 67]. However, this setup is not practical when the space of interactions spans over a large area and the event involves many participants.

For capturing human social interaction in-the-wild (e.g., networking event and conference), two types of sensor networks are commonly used for multimodal data acquisition: camera network [45, 46, 68] and wearable sensors network [69]. Not only does each network have to achieve intramodal synchronization, both networks need to be cross-modally synchronized in order to provide faithful data input to downstream multimodal machine learning models. Past works involved using post-processing or event-driven synchronization based on a subset of anchor frames [70, 71]. Cabrera-Quiros et al. [41] utilized a gossiping synchronization network approach [62] that guarantees global timestamps for wearable sensors to be accurate up to 1 second, but the synchronization with video data was also done manually.

Having identified the sensing challenges, new solutions for distributed and scalable multi-sensor data acquisition are needed. The proposed approach is to propagate a common time reference based on Network Time Protocol (NTP) to edge devices (i.e., wearable sensors and cameras) during acquisition. The NTP signal is converted to Linear Time Code (LTC) for the cameras and is converted to UNIX time for the wearable sensors. The crossmodal synchronization is achieved while preserving the existing modality-specific timekeeping. This approach not only alleviates manual or error-prone event based post-processing, but also ensures synchronization across modalities within milliseconds range. It is also scalable, reproducible, and cost-effective, which are also key requirements to increase the adoptability of this approach.

ANNOTATION ACQUISITION

The quality of collected annotations of behavioral cues depends on the nature of the data and annotation strategy. The types of data are consistently multimodal in this setup, and therefore, the annotation method needs to be adapted accordingly. In previous works such as MatchNMingle, bounding boxes and social actions are annotated using the Vatic tool [72] where annotators are presented with images at 1 Hz frequency and the intermediate samples are interpolated. This thesis uses annotations such as head and body orientations, speaking status, and conversation group membership. While there are design choices in speaking status annotations (visual, audio, or audiovisual) and conversation group membership (based on definition of F-formation), they are not the focus for the annotation acquisition aspect in this thesis. I discuss more specifically the annotations for head and body orientation which are the emphasis of Chapter 2 and 3.

For the augmentation of MatchNMingle, the head and body orientations were annotated via an adaptation of LabelMe [73], where annotators are presented with images for labelling keypoints. Specifically, we labelled head and body orientations for identifying upper body skeletal keypoints including left and right shoulder, center of the head, and the tip of

the nose (Chapter 2). With these keypoints, head and body orientations can be derived by taking the vector orientation from the center of the head to the nose, and taking the perpendicular orientation between left and right shoulder vector, respectively.

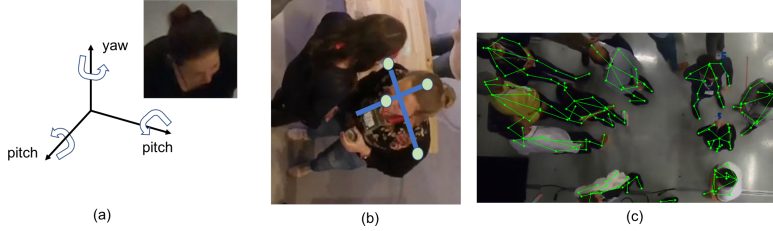


Figure 1.5: Data annotation strategies. (a) shows that from the overhead view, yaw axis corresponds the head movement panning in the scene. (b) shows annotation of upper body keypoints and face direction for extracting head and body orientations (from shoulders). (c) shows full body skeletal keypoints that provide more granular information than just upper body keypoints.

With this annotation strategy, the recent covfee tool was built to address the lack of temporal continuity in annotations [65] (used in Chapter 6). It seamlessly allows for keypoint annotation of videos with high frequency. Covfee reduces annotation time and achieves better inter-rater agreement [65]. Despite Covfee taking a great step in annotations for social signal processing, obtaining labels from various datasets is still time-consuming and expensive. Importantly, annotations can be subjective and noisy, which motivate modeling approaches that can account for these inevitable data characteristics.

SENSOR DEVELOPMENT

Existing wearable sensors that have been designed as conference smart-badges include Sociometric Badges, Chalcedonies, Rhythm Badges, etc. Rhythm Badges [63] possess the most enriched sensing capabilities with accelerometer, microphone, and Bluetooth proximity sensor. An extension of the accelerometer to IMU provides possibilities for derive orientations, as opposed to just body movement measurement. By including additional microphones, a microphone array could be used to estimate direction of arrival by leveraging phase information in signals from the spatially separated sensors. The recently developed Midge provides these extensions and have been deployed to enable future research in using these additional sensing capabilities.

1.3.2 MODELING APPROACHES

Further unpacking the modeling-related works that this thesis covers, which includes head and body orientation estimation, head orientation estimation in conversation groups, and conversation group detection in social interaction scenes, suggests that an interpretation of modeling social behavioral cues, which falls into three scales: (i) individual-level, (ii) group-level, and (iii) scene-level.

The fundamental component of a conversation group is individual members. Head and body orientations are naturally associated with properties of an individual (e.g., ranging from height and size of an individual which determines their body inertia and affects magnitude of their movement, to personality traits leading to more (e.g., extraversion)

or less expression and movement). On the other hand, head and body orientations of an individual could also be affected by group-level dynamics. Phenomena such as turn-taking, synchrony, proxemics, etc. and concepts such as dominance, deception, etc. are mostly defined in a group (>1 person) context as it requires social interaction in the first place. There is significant motivation to model head and body orientations on a group-level accounting for intragroup dynamics, as they are part of the behavioral cues that ultimately lead to social signals between interactants.

In addition to the intragroup dynamics, interaction scenes also feature intergroup dynamics when multiple groups are co-present. In complex conversational scenes, conversation groups are informal, but there is not restriction on them being interest, functional, or task, -focused as long as individual share the common activities. A common framework on group development, breaking the process into five stages, was proposed by Bruce Tuckman [74]. According to Tuckman's theory, the stages include: forming, storming, norming, performing, and adjourning. While the groups stabilize in storming, norming, and performing in a mostly self-regulating fashion, forming and adjourning of groups are directly related to scene level dynamics, as it is usually characterized by some confusion and uncertainty when individuals freely move around the scene or event. For the task of detecting conversational groups which has traditionally been done from a bottom-up perspective (constructing affinity matrix through pairwise affinities), it is also worthy to explicitly model the surrounding context, which is especially important in states where groups undergo many change.

1.4 CONTRIBUTIONS

In light of the standing challenges, this thesis tackles a subset of those, by focusing on novel methods to model human behavior in free-standing conversation groups in complex conversation scenes, and sourcing a high-fidelity in-the-wild interaction dataset with accurate multimodal synchronization. All methods-oriented studies in this thesis use multimodal data collected in-the-wild and account for the social context. All data-oriented studies in this thesis have been developed specifically to capture human interaction data in-the-wild. The intention is to develop methods based on data that are as ecologically-valid and close to real-life as possible.

The organization and contribution of the thesis is as follows:

- Chapter 2 presents a novel multimodal approach for joint head orientation estimation in a conversation group by leveraging the proxemics and dynamics within a social group, and the methods show to have some generalizable capability when applied to an unseen social interaction scene. The contribution also includes a large-scale upper keypoints and orientation annotation augmentation towards an existing dataset, MatchNMingle [41].
- Chapter 3 focuses on scene understanding and presents a conversational group detection method for an interaction scene using temporal based inputs such as positions, head, and body orientations. Approaching the task in two stages (i) affinity prediction and (ii) group detection via clustering results in an interpretation of the intermediate continuous affinity values based on the past, rather than only binary

group membership from direct clustering. Furthermore, this contribution includes a forecasting method that predicts future affinity values and conversation groups that may be useful in studying group and scene evolution.

- Chapter 4 introduces a transductive method developed for joint estimation of head and body orientation using multimodal data when annotations are scarce. The method takes advantage of the physical priors and inherent nature of the social interaction scene as additional information to elevated side-view videos and wearable sensor signals. It results in decent head and body orientation estimation especially in the low-data regime.
- Chapter 5 presents a hardware synchronization of multisensor multimodal data during acquisition based on network time protocol. It eliminates the time-consuming post-hoc manual alignment and/or ad-hoc event based synchronization which may not generalize to wearable sensors.
- Chapter 6 describes an instantiation of best practices in collecting human interaction data, and offers a novel high-fidelity dataset collected during a real-life professional networking event. The fine-granular annotations and high-quality of multimodal data closes the gap for future development of automated methods.

Respective publication information is indicated in each chapter title page.

2

MULTIMODAL JOINT HEAD ORIENTATION ESTIMATION IN INTERACTING GROUPS VIA PROXEMICS AND INTERACTION DYNAMICS

This chapter is published as:

Stephanie Tan, David MJ Tax, and Hayley 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 (2021): 1-22.

2.1 INTRODUCTION

Social interactions are one of the most fundamental aspects of human behaviors. If machines were able to understand these behavioral patterns, they would have more potential to perceive, interpret, predict and even influence behavior. In the context of conversing groups, the behavioral patterns of a conversant are not isolated but rather coupled with the behaviors of other participants. In this paper, we focus on the automated analysis of head orientation, which is an important cue for social attention, particularly in the absence of eye gaze measurements. In the context of conversations, humans orient their heads based on the flow of a conversation, such as re-orientating their head when there is a change of speakers [75], when a new participant joins the conversation [76], or when there is a change of head and body orientation of other participants [77]. Head orientations, as a proxy for direction of attention [16, 78], can be reflective of the participant's mental processes and therefore, the interaction quality which is valuable information towards social intelligence [79]. In order to obtain such information, a deeper understanding of the interaction between conversing people is required. These interplays include the evolving proxemics (relative positions and orientations) and conversation dynamics (e.g., turn taking behavior, spontaneous responses to a speaker or listener) of an interacting group. This paper focuses on the automatic estimation of head orientation in relation to these complex phenomena, specifically in crowded mingling (in-the-wild) social scenes.

We first establish the importance of the interaction context that we are interested in, which is different from the ones in previous studies and their interpretations. While human interactions in focused settings [19, 20, 38], e.g., seated meetings (Figure 2.1(a)), have been studied extensively, a closer analysis of complex conversational scenes [80], e.g., networking events or cocktail parties [6] (Figure 2.1(b) and 2.1(c)), is more challenging. We differentiate interacting groups in complex conversational scenes from free-standing conversations group (FCG) which have been studied in the past [6]. FCGs form spontaneously as soon as people gather in close vicinity to sustain a common space and they are motivated by proxemics alone. On the other hand, an interacting group in complex conversational scenes has another layer of complexity when modelling its members' behaviors. A group could contain multiple conversation (sub)groups and thus more varied interaction dynamics, though still sharing the common physical space [81]. Due to the noisiness and unstructuredness of in-the-wild settings, the underlying conversational and behavioral patterns are different from those of seated meetings and other formal interactions, and further, are not attributed to only proxemics as with FCGs. In addition, social scientists have shown that interacting people tend to exhibit movement coordination [12, 24]. Body movement could serve as informative cues towards head orientation estimation due to anatomical constraints, and also account for the possibility of group level phenomenon such as movement mimicry and synchrony [82]. We argue that head orientations should be studied in consideration of the interaction context, which is coupled with the underlying interaction dynamics, represented by changes in proxemics, speech, and body movement.

Previous methods do not explicitly address the dynamic context for interacting groups in complex conversational scenes. Many existing methods are designed for head orientation estimation specifically for meeting analysis (e.g., strapping sensors onto participants' head [19] (Figure 2.1(a)), OpenFace [83], etc., among other methods that use audio data [84, 85]). For our setting (Figure 2.1(b) and Figure 2.1(c)), adopting a wired connection for direct

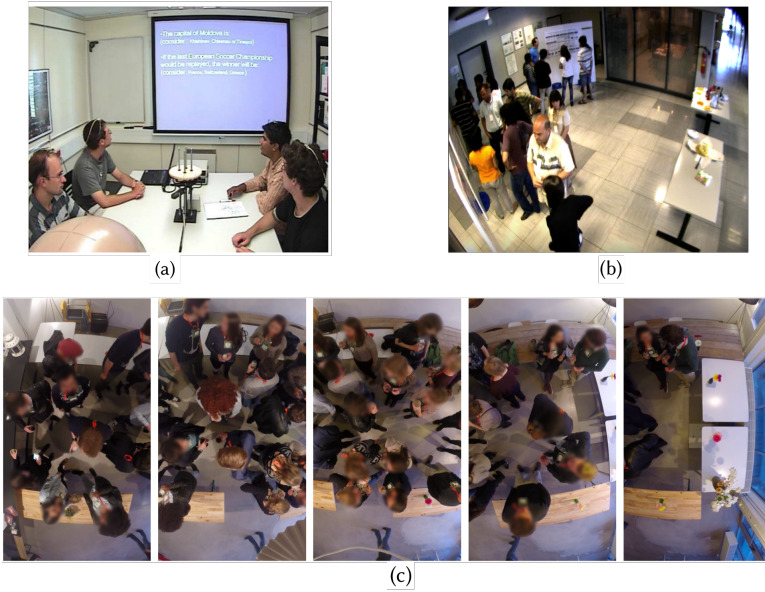


Figure 2.1: Examples of types of scenes where people interact with each other in groups. (a)[19] and (b)[45] show a meeting and poster session in which participants have shared targets of attention (screen, poster board, etc.). In (c)[41], the unstructuredness makes studying the social interactions hard.

measurement of subjects that span a large physical space or placing close-up (third or first person) cameras for facial images is not feasible and also undesirable. Other appearance-driven methods (e.g., [29, 50, 86]) for head orientation estimations have been developed for these complex conversational social scenes. They typically rely on a small number of overhead or elevated cameras. While these camera perspectives capture the whole ground plane, head orientation estimation remains challenging [50] because of – (i) low-resolution images, (ii) a high degree of self-occlusion and occlusion by other people, and (iii) missing informative features such as facial attributes. In some of these works, the dynamics context of interacting people was utilized to improve head orientation estimation, but it was motivated from only the proxemics aspect. Despite this additional consideration of context, occlusions are generally the main factor causing poor estimations [50, 87]. The Panoptic studio [18] solution more explicitly addressed challenges related to occlusions but is only realized by using hundreds of cameras in a specialized infrastructure which would be hard to replicate in real life. For these reasons, modeling head orientation estimation using appearance features for crowded settings in-the-wild remain as open question. Multimodal methods [28, 88] have shown promise towards improved head orientation estimation by utilizing additional modalities.

In this light, we propose a multimodal approach towards head orientation estimation method that takes advantage of a small number of room mounted overhead cameras and wearable sensors. In particular, we consider a single wearable sensor worn around the neck, like a smart ID badge, which records body inertial motion, speaking status, and proximity, akin to the ones used in [41] and [63]. The multimodal solution aligns well

with estimating head orientations by considering the dynamics in proxemics, speech and body movement, as we motivated. We show that it is possible to accurately model head orientations for in-the-wild conversation scenes solely based on a novel integration of proxemics (relative position and orientations) and interaction dynamics (speaking status and tri-axial acceleration from a single accelerometer hung around the neck) and without relying on vision-based appearance data.

Further, we argue that since head orientations are a proxy for attention and could be dynamically changing in more chaotic settings, head orientation estimation should be formulated as a regression task. Existing orientation estimation methods (e.g., [28, 29, 50, 86, 88]) treat the task as 8-class classification problem. This discretization may result in a loss of fine-grained information. Additionally, a classification setting assumes that each class is independent from one another, which is an incorrect assumption for estimating head orientations, which are intrinsically continuous.

To this end, we adopt a long short-term memory (LSTM) based approach wherein we take signals of different modalities over time to estimate continuous head orientations. Different modalities provide complementary information and lead to a multivariate time series modeling problem. To account for the group context using a more rigorous concept based on social science, we consider interacting groups as *F-formations* [12], in which participants collaborate to maintain an interaction space by establishing spatial and orientational relationship. The inputs of our system are temporally aligned sequences of position, body orientation, speaking status and tri-axial acceleration of *all* members of an F-formation; the outputs of our system are the head orientations (i.e., continuous angles) of the same group of people at the last timestep of the aforementioned sequences.

More importantly, the underlying intuition is that humans orient their heads based on the behavior of other people in their conversation group. This motivates us to build a model which accounts for the dynamic interaction of members in the same social group when estimating their head orientations. Our results show that (1) estimating head orientations considering the group context is better than considering individuals only; (2) temporal context is more advantageous than only estimating at a temporal snapshot; (3) including tri-axial acceleration and speaking statuses (indicative of body movements and conversation dynamics, respectively) contributes positively to model performance; (4) the model generalizes well to estimating head orientations in groups of different sizes and also unseen data; and (5) training with continuous labels results in lower errors than with discretized labels (head orientation angles binned into pre-defined sectors).

To this end, we list our contribution as follows:

1. a novel feature set for estimating head orientations in crowded settings,
2. a pooling module that explicitly integrates information of all interaction partners to jointly model the dependence between each person in an interaction segment,
3. a detailed analysis of head orientation estimation performance with respect to different methods, generalization to unseen data, and sensitivity to different modalities,
4. an experimental validation of formulating head orientation estimation as a regression task, as opposed to an 8-class classification task,

5. and lastly, a large-scale annotated data resource containing upper body keypoints (shoulders and head), head and body orientations, and F-formation group membership based on the MatchNMingle dataset [41].

In Section 2.2, we review the related works on head orientation estimation. Section 2.3 presents the details of our approach, methodology, and implementation. Section 2.4 discusses the dataset and our design choice based on it. Section 2.5 presents the relevant results and analyses. We close by discussing future work and conclusion in Sections 2.6 and 2.7.

2.2 RELATED WORK

In this section, we introduce relevant previous works on orientation recoveries and estimation, using wearable sensors and cameras. Human head orientation estimation, as a subtopic in general orientation estimation, has attracted much attention from researchers working on human-computer interaction, pose estimation, and other related topics. In Section 2.2.1, we summarize related works in orientation measurement and estimation using wearable sensors. In Section 2.2.2, we present some representative works on head orientation estimation using room mounted cameras, particularly in social settings similar to ones that we are interested in.

2.2.1 WEARABLE SENSOR BASED ORIENTATION MEASUREMENT AND ESTIMATION

IMUs containing accelerometer, gyroscope, and magnetometer can be used to obtain orientation estimations. Kok et al. [89] provide a comprehensive overview of methods that integrate raw sensor outputs into orientation recoveries and estimations. However, IMUs suffer from heading drift during continuous operation and the accuracy of the resulting angle estimations could be compromised, especially in indoor scenarios where magnetometer measures are noisy [90]. Many motion capture systems feature large number of IMUs attached to custom made suits to capture positions and orientations in everyday surroundings [91]. However, these approaches do not align with the reconstruction using video and do not address the drifting issue [92]. To compensate for the drift, multiple cameras and IMUs are combined to estimate orientations [93, 94]. Most of these previous works focus on human pose estimation and takes advantage of placing of sensors on different limbs and optimizing via consistency in joints and segments [92, 93]. If we are only interested in head orientation estimation, multiple sensors would have to be placed on the head, in conjunction with setting up other cameras. This quickly becomes infeasible in in-the-wild settings. For lab-like settings, the IDIAP head pose dataset [19] was collected using a 3D location and orientation tracker that supported tracking four sensors strapped onto the participant heads using the Flock-of-Birds magnetic sensing technology at a 50Hz sampling rate, in order to enable the study of higher-order behavior such as visual attention via head pose. This particular solution only supports up to four sensors and requires a wired connection, which makes it difficult to scale up and deploy for in-the-wild social interactions. More recent IMU-based solutions that are more mobile, such as the earables

[95], have been proposed. However, their study has shown that great variability exists in how people wear the sensor in the ear (i.e., different angle/orientations), which further complicates estimating head pose between different subjects. While it is also possible to estimate head orientation using only wireless wearable sensors (e.g., smart ear pieces and head bands) [96, 97], these technology are still in nascent stages and are not able to estimate orientations accurately. A recent technology based on near-infrared sensing utilizes pairwise light sensing to infer the incident angle and distance between two sensors [98]. However, there are collisions in transmission when there are multiple (>2) sensors involved. Using this approach, sensor reading is dropped when there are detected collisions, hence reducing the detection frequency (once in five seconds) to an even lower and variable rate. This technology is also not readily extendable to detecting head orientations and studying human behaviors on a finer temporal granularity in crowded scenarios. Aside from the technical limitations, putting these sensors on people's head to measure orientations raises concerns of social acceptability [99]. Such practice would violate the ecological validity [100] of the observations of human behaviors. On the other hand, the use of wearable badges around the neck is less intrusive [101] and could be used in settings where devices such as Google Glass or cell phones are considered impolite or forbidden.

2.2.2 ROOM MOUNTED CAMERA-BASED HEAD ORIENTATION ESTIMATION

Image-based head orientation estimation is represented by a large body of literature [87, 102–110]. When the face features are available, these methods are able to predict head orientation in all 3 axes (roll, pitch, and yaw), more generally called head pose estimation. Note the majority of these works in this topic do not consider room-mounted cameras but using cameras (e.g., webcams) that allow for capturing facial images. Hence, they are not directly applicable in the surveillance setting where subjects are far away and images are lower-resolution. For the rest of the discussion in this section, we focus on works that use camera inputs acquired from an angled height. In the setting of predicting poses of pedestrians, previous works [49, 111] take advantage of the motion prior from trajectories and/or the coupled body orientation to compensate for the lack of high-quality visual input. However, in a crowded scenario where human subjects are mostly static and occlusions are frequent, these approaches become ineffective. Ricci et al. [50] explicitly tackle the occlusion problem when jointly estimating head and body orientations. The method requires determination of the level of occlusion based on targets' feet and head locations obtained from tracking, which is feasible in elevated side-views but remains challenging in overhead views where head and body crops overlap. Moreover, head orientation estimation under these adverse settings has usually been formulated as a classification problem. Typically, orientations are divided into 8 classes, with a bin size of 45° [28, 88] since most available datasets for head orientation estimation in low resolution only include discretized class labels. Even though some works such as by Yan et al. [87] reported estimation errors in degrees, they converted the estimates from discretized class labels according to the center of the bins and the results do not imply that the model was trained with respect to continuous outputs. With recent advances in deep learning methods and increased efficacy of convolution neural networks, Prokudin et al. [112] have shown high accuracy on low resolution head images. However, for the scope of this paper, our contribution is

orthogonal to comparing against state-of-the-art vision based methods, as we wish to show the efficacy of a new set of inputs.

We point out that some existing vision-based methods have taken the social context into account when estimating head orientations, such as [28, 29, 50, 86]. They provide solutions for the joint estimation of group membership and head and body orientations. However, we differentiate our paper from these works as we explicitly model for the interplay between subjects, with the consideration of evolving proxemics and dynamics captured by speech behavior and body movement. In that regard, our work is most similar to that of Otsuka et al. [113] where a multimodal and multiparty fusion method was proposed to estimate visual focus of attention, albeit a different task from ours. Their experiments showed that a group based model outperformed individual based model in certain cases (a promising clue that inspired our paper). However, this work, among others such as [19, 20], is constrained to estimating visual focus of attention in a focused and structured meeting scenario, which lacks changes in proxemics and body movement. Instead, our work focuses more on in-the-wild settings, with inputs that could be obtained without using close-up cameras.

2.3 METHODOLOGY

2.3.1 APPROACH

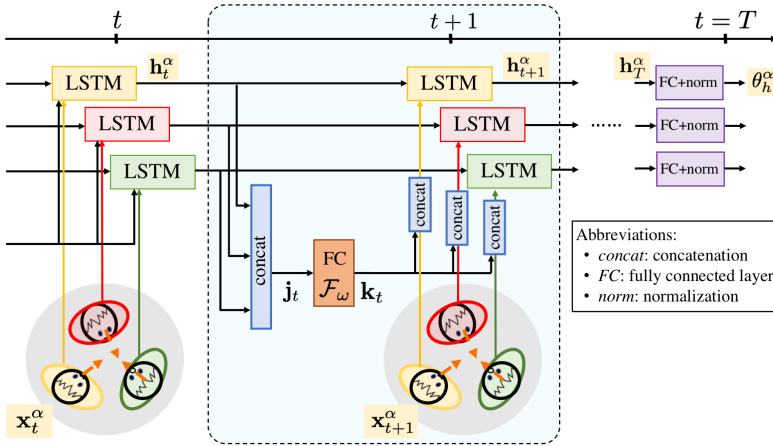


Figure 2.2: Graphical illustration of our approach for head orientation estimation in the group context.

For a given group of size G , let $\mathbf{x}_t^\alpha \in \mathbb{R}^N$ denote the feature vector for member $\alpha \in \mathcal{G}$ at sequence step $t \in \{1, \dots, T\}$, where $\mathcal{G} = \{1, \dots, G\}$ is the set of members in the group, T is the sequence length, and N is the number of features. The feature vector \mathbf{x}_t^α is a concatenation of the following:

- speaking status of member α ,
- three-channel (raw) signal from tri-axial accelerometer of member α ,
- body orientation of member α ,

- positions of all the members $\{\beta \in \mathcal{G} : \beta \neq \alpha\}$ relative to the member α in polar coordinates (radial distance and angular orientation).

To capture the group context, we adopt a shared LSTM-network based approach. Here, we exemplify the recurrent step of the proposed model from t to $(t+1)$ which is schematically summarized in Figure 2.2. Let $\mathbf{h}_t^\alpha \in \mathbb{R}^H$ denote the hidden states associated with member α at sequence step t , where H is the number of hidden states (chosen as a hyperparameter). The hidden states at $t = 1$ are initialized as $\mathbf{h}_1^\alpha = \mathbf{0}$. At any step $t < T$, the hidden states from the current step of all the members in the group are concatenated into a hidden representation $\mathbf{j}_t \in \mathbb{R}^{G \times H}$ as

$$\mathbf{j}_t = [\mathbf{h}_t^1; \mathbf{h}_t^2; \dots; \mathbf{h}_t^G], \quad (2.1)$$

where $[\cdot; \cdot]$ denotes concatenation. \mathbf{j}_t is then mapped into a lower dimension K ($K < G \times H$) using a linear layer \mathcal{F}_ω with rectified linear unit (ReLU) as activation function to obtain $\mathbf{k}_t \in \mathbb{R}^K$ as

$$\mathbf{k}_t = \text{ReLU}(\mathcal{F}_\omega(\mathbf{j}_t)), \quad (2.2)$$

where ω denotes the set of weights and biases of the linear layer. This step performs *context pooling*, i.e., the information stored in the hidden states associated with each individual are combined together to obtain a group-level context represented by \mathbf{k}_t . Thereafter, for each member α , the respective hidden state \mathbf{h}_t^α as well as the concatenation of \mathbf{x}_t^α (individual member's input) and \mathbf{k}_t (group-level context) are passed to an LSTM cell \mathcal{L}_τ (parameterized by the set τ) to obtain the output hidden states

$$\mathbf{h}_{t+1}^\alpha = \mathcal{L}_\tau([\mathbf{x}_{t+1}^\alpha; \mathbf{k}_t], \mathbf{h}_t^\alpha). \quad (2.3)$$

The LSTM operation \mathcal{L}_τ is described by the following series of transformations

$$\begin{aligned} \mathbf{f}_{t+1}^\alpha &= \sigma \left(\mathcal{W}_{\xi_f}([\mathbf{x}_{t+1}^\alpha; \mathbf{k}_t; \mathbf{h}_t^\alpha]) \right) && \text{(forget gate's activation vector)} \\ \mathbf{i}_{t+1}^\alpha &= \sigma \left(\mathcal{W}_{\xi_i}([\mathbf{x}_{t+1}^\alpha; \mathbf{k}_t; \mathbf{h}_t^\alpha]) \right) && \text{(input gate's activation vector)} \\ \mathbf{o}_{t+1}^\alpha &= \sigma \left(\mathcal{W}_{\xi_o}([\mathbf{x}_{t+1}^\alpha; \mathbf{k}_t; \mathbf{h}_t^\alpha]) \right) && \text{(output gate's activation vector)} \\ \tilde{\mathbf{c}}_{t+1}^\alpha &= \tanh \left(\mathcal{W}_{\xi_c}([\mathbf{x}_{t+1}^\alpha; \mathbf{k}_t; \mathbf{h}_t^\alpha]) \right) && \text{(cell input activation vector)} \\ \mathbf{c}_{t+1}^\alpha &= \mathbf{f}_{t+1}^\alpha \odot \mathbf{c}_t^\alpha + \mathbf{i}_{t+1}^\alpha \odot \tilde{\mathbf{c}}_{t+1}^\alpha && \text{(cell state vector)} \\ \mathbf{h}_{t+1}^\alpha &= \mathbf{o}_{t+1}^\alpha \odot \tanh(\mathbf{c}_{t+1}^\alpha) && \text{(output hidden state vector),} \end{aligned} \quad (2.4)$$

where σ is sigmoid activation, \odot denotes the Hadamard product, and \mathbf{c}_t^α and \mathbf{c}_{t+1}^α denote the cell state at t and $(t+1)$, respectively. $\mathcal{W}_{(\cdot)}$ denotes a linear layer with parameters indicated in the subscript. The trainable parameters are contained in the set $\tau = \{\xi_f, \xi_i, \xi_o, \xi_c\}$. Importantly, the LSTM parameters τ are shared among all the members of the group.

After recursing through the LSTM cell (3.3) upto $t = T$, the hidden states \mathbf{h}_T^α are passed through another linear layer \mathcal{K}_κ (parameterized by κ),

$$\mathbf{q}^\alpha = \mathcal{K}_\kappa(\mathbf{h}_T^\alpha), \quad (2.5)$$

where $\mathbf{q}^\alpha \in \mathbb{R}^2$; followed by a normalization layer

$$\mathbf{y}^\alpha = \frac{\mathbf{q}^\alpha}{\|\mathbf{q}^\alpha\|}. \quad (2.6)$$

Similar to the LSTM cell, the parameters κ are shared among all the members. Note that this step is only performed after all the timesteps are processed. Due to the normalization, $\mathbf{y}^\alpha = (y_1, y_2)^\top$ may be interpreted as cosine and sine of the head orientation prediction θ_h^α . To this end, we obtain θ_h^α from \mathbf{y}^α as

$$\theta_h^\alpha = \text{atan2}(y_2, y_1), \quad (2.7)$$

where $\text{atan2} : \mathbb{R} \times \mathbb{R} \rightarrow (-\pi, \pi]$ is the 2-argument arctangent.

To train the model, we use the cosine similarity loss function [112, 114] summed up over all the members in \mathcal{G}

$$\ell = \sum_{\alpha=1}^G (1 - \mathbf{y}^\alpha \cdot (\cos \theta_{h,GT}^\alpha, \sin \theta_{h,GT}^\alpha)^\top), \quad (2.8)$$

where $\theta_{h,GT}^\alpha$ represents the ground truth head orientation. Considering a training set of n such sequences, we minimize the loss over all the sequences to optimize the model parameters

$$\omega^*, \tau^*, \kappa^* = \underset{\omega, \tau, \kappa}{\text{argmin}} \sum_{i=1}^n \ell_i, \quad (2.9)$$

where ℓ_i is the loss associated for the i^{th} sequence in the dataset. Note that the group size (G) may vary within the dataset.

2.3.2 IMPLEMENTATION DETAILS

The feature set is obtained through manual annotations from the overhead camera view, except the tri-axial acceleration which are obtained from the wearable sensor directly. As described earlier, we use these annotations, where applicable, as proxy to avoid confounding sources of error in our inputs. The comparison of different methods to acquire automated inputs of better quality is out of scope for this paper. Speaking status is a binary value where 0 and 1 denote "not speaking" and "speaking", respectively. The body orientation is given by the angular direction of the person's body in $(-\pi, \pi]$. The relative positions of the group members are given by the radial distance (measured in pixels) and angular orientation (in $(-\pi, \pi]$). In practice, the positional angular orientations are computed in reference to the circular mean of the same over all the group members (except the member in question), i.e., the latter serves as the zero-degree reference. This removes the discontinuous jump in angles as they wrap around $(-\pi, \pi]$ and removes sensitivity to the group location in the scene as well as other group-specific attributes. Body orientations of all group members are also corrected by the same zero-degree reference. Continuous feature values (acceleration, body orientation, relative distance and orientation) are normalized to the range $[0, 1]$ by min-max scaling. Finally, we note that group membership is considered as pre-determined and assigned based on Kendon's definition of F-formation [12].

We perform a three-fold split of the available groups (not the sequences as described below) for cross-validation such that groups in the validation splits do not appear in training. Given the lifetime of a group (i.e., the duration in which no new member joins and no

The circular mean $\bar{\alpha}$ of a set of angles $\{\alpha_1, \dots, \alpha_n\}$ is computed as the arctangent of mean of sine and cosine of all the angles, i.e., $\bar{\alpha} = \text{atan2}(\frac{1}{n} \sum_{j=1}^n \sin \alpha_j, \frac{1}{n} \sum_{j=1}^n \cos \alpha_j)$, where $\text{atan2} : \mathbb{R} \times \mathbb{R} \rightarrow (-\pi, \pi]$ is the 2-argument arctangent.

existing member leaves), we generate several sequences using a sliding window of stride equal to one. The sequences are of length $T = 10$, which is a design choice that we further justify in Section 2.4. The number of model outputs is catered to the biggest group size of $G = 7$ members (observed over 300 distinct groups in 90 minutes of video recordings in the MatchNMingle dataset [41]). The members are arranged in a random order. The training dataset is augmented by shuffling the member order to achieve better results on the validation set. More details on the data augmentation procedure can be found in the auxiliary materials. Only the outputs corresponding to the relevant group size are considered for evaluation, since the group size is known a priori. Missing feature values in smaller groups are padded with a constant value which we set to -2, which suffices to inform the neural network for missing values rather than noise in the inputs. We find that choosing different padded values does not affect the model performance.

The performance is reported for the following hyperparameters which was obtained through a grid-search on a subset (10 min) of the MatchNMingle data. The dimension of the LSTM hidden states is set to $H = 20$. The output from the context pooling module is of dimension $K = 32$. ADAM optimizer is used to train the model for 100 epochs with a learning rate of 10^{-4} and batch size of 16.

To intuitively understand the model performance, we choose the root mean squared error RMSE = $\sqrt{\frac{1}{n} \sum_{j=1}^n \Delta_j^2}$ over n test samples, where we define the angular difference Δ as

$$\Delta(\theta_1, \theta_2) = \min(|\theta_1 - \theta_2|, 2\pi - |\theta_1 - \theta_2|), \quad \text{with } \theta_1, \theta_2 \in (-\pi, \pi], \quad (2.10)$$

since (head-orientation) angles wrap around with period 2π .

2.3.3 BASELINE METHODS

We compare the proposed method to three baseline methods. We first consider a rule-based method, which is engineered based on knowledge of patterns in conversation dynamics. Two other methods representing controlled settings of our proposed method are considered to illustrate the effects of the temporal context and individual vs. group based inputs.

The rule-based method of head orientation estimation is inspired by previous works [19, 21, 38, 75] for the task of estimating visual focus of attention (VfOA) in meeting scenarios. They utilize Bayesian methods (e.g., dynamic Bayesian networks) to model the roles of contextual information such as head pose, speaking status, conversation structure, etc. These models are built on domain knowledge, which is expressed in a causal structure relating different variables. Though these methods focus on a different task, there is high-level similarity to the proposed method, which is to include multimodal, multiparty, and contextual information. In the spirit of designing a model that uses expert knowledge of the phenomena in question, we devise the simplified rule-based method to capture a specific type of dynamics; i.e., listeners tend to orient their head towards the speaker. At any given time, a listener's head orientation is given by the orientation of the positional vector from the listener to the speaker. If there are two or more speakers in the group, then the circular mean is computed over the respective orientations. A speaker's head orientation is given by the circular mean of the orientation of the positional vectors from the speaker to each of the listener. The speakers and listeners are identified based on the speaking status.

Without including the temporal information, we propose a frame-based method as follows. For a given member in the group, we use a multi-layer perceptron (MLP) that takes as input the accelerometer signals of the member in question, the relative positions (radial distance and angular orientation) of all the other members, and the speaking status and body orientation of all the members in the group. The MLP does not consider temporal information and only predicts on a frame-wise basis.

To include the temporal information, we propose a sequence-based method which uses an LSTM-based network, acting on a sequence of the same set of inputs as the frame-based method. The sequence length is also chosen to be $T = 10$ to match with the design choice of the proposed model. This model is a simplified version of the proposed model which does not pool the hidden states of other members of the group.

The frame-based and sequence-based methods are both considered as *individual models* because they use inputs arranged from an individual group member's perspective. In our proposed method, which we consider as *group model*, we *jointly* estimate the head orientations of all members in an interacting group by considering the relative information between all possible pairs of individuals and context pooling the hidden states between subsequent steps of LSTM. This conceptual difference is illustrated in Figure 2.3. Each colored area indicates the area of influence of an individual. The hexagon in Figure 2.3(b) delineates an interaction space containing three individuals interacting with one another, as an example. Figures 2.3(a) and 2.3(c) represent the frame-based and sequence-based method, respectively, and the estimation only concerns the bottom individual (denoted in pink). Figure 2.3(a) shows that the frame-based method only considers inputs at a single time step, whereas Figure 2.3(c) shows the inputs progressing in time, which is reflected by the sequential inputs to the sequence-based method. Although considering the presence and the information of other two members (top) in the same interaction space, these methods do not take into account that the interaction space is shaped by all 3 individuals (Figure 2.3(b)). On the contrary, this factor is incorporated in the proposed group-based model, where the hidden states of all members are pooled into a unified representation at each timestep to track how members influence each others' head orientations in an evolving interaction (Figure 2.3(d)).

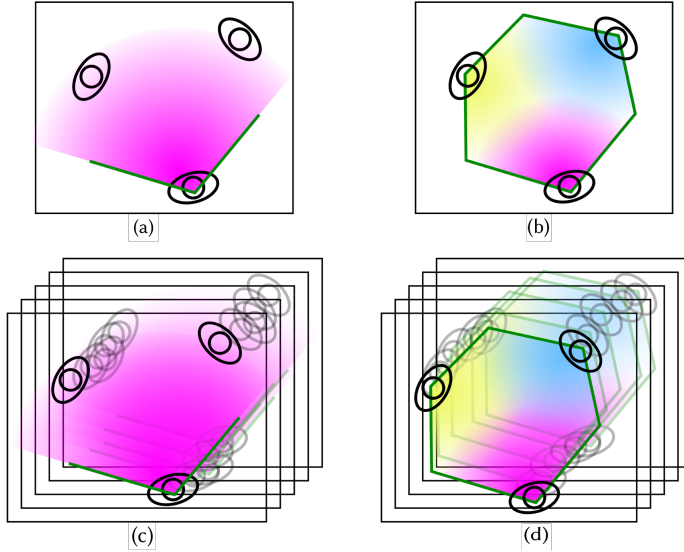


Figure 2.3: Conceptual visualization of the motivation behind different methods. (a) The inputs and output are designed from an individual's perspective in the group. The frame-based method is designed for this scenario. (b) The inputs and outputs are designed from a group perspective. (c) A sequential version of (a) where the change of interaction space in time is modeled using the sequence-based method (d). The proposed method, which is group-based, considers the change of the joint interaction space given that individuals influence each other during the course of an interaction.

2.4 HEAD ORIENTATIONS IN COMPLEX SOCIAL SCENES: A CASE STUDY

2.4.1 DATASET

We develop and test our model on the MatchNMingle dataset [41] which is one of the largest multimodal datasets capturing human interactions in-the-wild. The dataset was recorded in casual networking events over 3 days and includes 90 minutes of unscripted and free interactions among 92 unique participants. There are 32, 30, and 30 participants for each day of the event, respectively, for 30 minutes each day. Subject positions, speaking status, body orientation, head orientation, and F-formation group membership were annotated by human annotators through visual perception of the overhead-surveillance video data at 1 Hz. The wearable sensor directly outputs raw tri-axial accelerometer data at 20 Hz. As a preprocessing step, we downsampled accelerometer data to 1 Hz by taking an average for each of the axis for each 1 second window. While using a higher frequency signal is understandably desired, human head rotation of 1 Hz and considered as a medium level of activity from a physiological point of view. Vigorous head motion activities, which are rare (head impulses), are in the range of 2.6 Hz [115]. Therefore summarizing the information from the tri-axial accelerometer into 1Hz should already capture predominant head behaviors.

This dataset is suitable for our studies because it contains a large number of people

Table 2.1: Comparison of different datasets available/used for head orientation estimation. Here, annotations pertain to the head orientations in the respective datasets.

Dataset	Context	No. of subjects	Length [minutes] × no. of segments	Modality	Annotation resolution [Hz]	Annotation method
IDIAP [19]	seated meeting	4; ×8 meetings	10; ×8 meetings	Video	N/A	FOB† sensor
TownCentre[110]	pedestrian	2200	22	Video	N/A	Human
CAVIAR[116]	pedestrian (scripted)	40	1; ×17 segments	Video	25	Human
CocktailParty [50] ‡	FCG	6	30	Video	N/A	Automatic
Coffeebreak[22]	FCG	14	2; ×2 segments	Video	N/A	Automatic
SALSA-Cocktail Party[45]	FCG	18	25	Video, Wearables	1/3	Human
MatchNMingle[41]	FCG	32, 30, 30 (3 events)	30; ×3 events††	Video, Wearables	1	Human

† Flock-of-Birds: head pose tracking using 3D magnetic sensors

‡ While more annotations of body and head orientations have been used in previous works [50, 86], only automated estimations of positions and head orientations are publicly available.

†† Due to occlusions near the edges of the camera field-of-view and design choices of strategically annotating before and after the lifetime of a group, the estimate is an approximation of the total number of annotations.

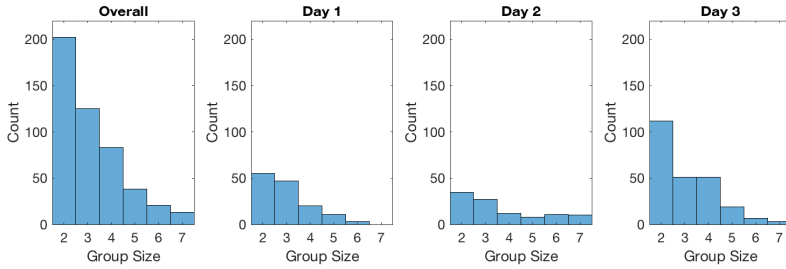


Figure 2.4: Distribution of group sizes in the MatchNMingle dataset.

forming a large number of different groups. Additionally, various social interaction data including speaking status as well as wearable sensor information (acceleration and proximity) are available. With the additional annotations of head and body orientations that we include as part of this paper’s contribution, the MatchNMingle dataset is the most fitting for our goals, as other datasets are 1) not situated in similar social settings, 2) of smaller scale and coarser temporal granularity, or 3) not as enriched in terms of data modalities. We compare the different available datasets for head orientation estimation in Table 2.1 to illustrate this point. The SALSA-Cocktail Party dataset is the most similar to the MatchNMingle. For completeness, we demonstrate the generalization of our model to unseen data from the SALSA dataset in Section 2.5.

The group size distribution of all the three days in the MatchNMingle data is shown in Figure 2.4. Smaller groups are more common than larger groups. Figure 2.5 shows that smaller groups have longer duration on average, though the variance is considerably high in most cases. We note that the group size and group duration distribution differ between the 3 days of data collection even though the social context is similar.

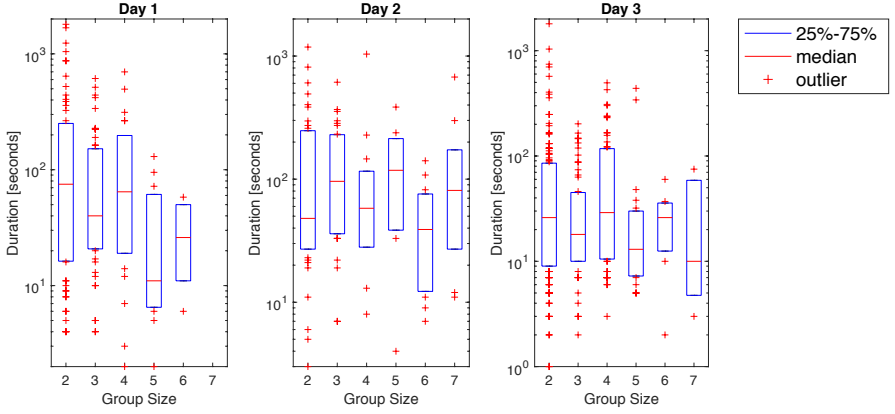


Figure 2.5: Distribution of group durations in the MatchNMingle dataset.

2.4.2 HEAD ORIENTATION ANALYSIS

Annotation. The head orientations are annotated by labeling head and shoulder keypoints. These annotations are performed by crowdsourcing workers. To quantify annotation discrepancy, we re-annotate a set of 2000, randomly sampled, data points out of a 10-min segment from the Day 1 recording of 32 participants. We found the difference in two sets of annotated head orientations to be $17 \pm 5^\circ$.

Justification for regression over classification. Head movements in relatively intimate social contexts (as opposed to pedestrian contexts) are fast changing. The argument in favor of regression is two-fold: (i) modeling orientations in a classification setting assumes classes/categories to be independent from one another, while angles are ordinal; (ii) head movement could be completely undetected if orientations are expressed in terms of classes. For the first argument, even if we discretize angles into classes, we can't easily model for the "closeness" of class 4 and 5 in a classification setting, for example. This lends naturally to a regression formulation. A potential argument in favor of classification applies when the data is so sparse that they could indeed be treated as independent classes. However, this is not the case in our data as we show that angles in our data span over the whole 360° range (see auxiliary materials). For the second argument, we illustrate using an example from the MatchNMingle dataset in Figure 2.6. We compare the head orientation time series of a subject to its discretized version with eight classes (as a 45° class bin-width is commonly assumed in previous works [28, 29, 50, 86, 88]). Notably, some segments of visible head turns of small magnitude are binned into the same class. The discretization results in a loss of valuable information that could be associated to related interaction events. This is especially relevant in big groups that span a large physical area; a small angular shift in head pose from one side of the group could indicate a shift in interacting partner on the far side. As the discretization over angles is arbitrary, there can be misleading fluctuations at the class boundaries. Since angles are continuous in nature and the limitation of low-resolution head images is no longer relevant in our case, modeling head orientations in a classification setting provides no definite advantages over regression. To this end, we show the comparison of classification vs. regression in Section

2.5.

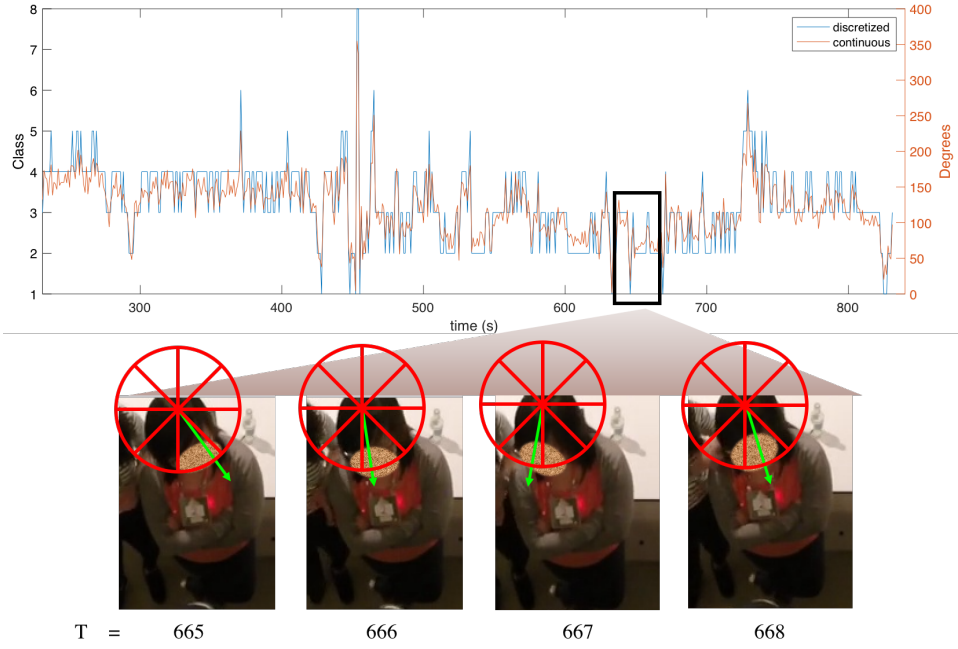


Figure 2.6: Illustrative example of how small but noticeable head turns could result in the same class bin, resulting in a loss of fine-grained information. From $T=665$ to $T=666$, the head turn with a relatively large magnitude resulted in the same bin. However, from $T=666$ to $T=667$, a head with a smaller magnitude resulted in different bins.

Sequence length. We make an informed design choice for the sequence length (T) in the LSTM by observing the speech behavior in the interactions. More specifically, we quantify the speech duration by measuring the length of segments when the speaking status is continuously equal to one. Without considering speech overlap and time delay in speaker transitions, we find the speech duration to be on average 4.8 seconds in the dataset. To ensure the possibility of including a speaker transition, we choose the sequence length for LSTM to be 10 seconds. At 1 Hz of sampling frequency of the feature set, this corresponds to $T = 10$.

2.5 RESULTS AND DISCUSSION

In this section, we present our experimental findings. A summary of the performance of the aforementioned baseline methods (see auxiliary materials for details) and our proposed method is shown in Table 2.2. In Section 2.5.1, we compare and analyze the performance of the proposed method and the baseline methods. For the proposed model, we discuss its generalization to unseen data, contributions and sensitivity of different combinations of inputs in Section 2.5.2-2.5.3. In Section 2.5.4, we compare our proposed method which is based on regression to an 8-class classification setting. In Section 2.5.5, we assess the model

Table 2.2: Summary of head orientation estimation performance of the proposed method and the baseline methods (introduced in Section 2.3.3). All methods are trained and evaluated with the same cross-validation splits.

Method	Type	Mean RMSE in θ_h (std. dev.) [$^\circ$]		
		Day 1	Day 2	Day 3
rule-based	static	47.7 (31.7)	44.1 (28.5)	54.4 (36.6)
frame-based	static	26.4 (18.7)	23.9 (15.4)	35.3 (25.7)
sequence-based	temporal	25.9 (20.0)	26.4 (16.6)	28.6 (17.5)
(proposed) group-based	temporal	22.7 (14.0)	22.0 (12.6)	25.0 (15.9)

sensitivity with respect to body orientations. And lastly in Section 2.5.6, we show how our model handles dynamic moments in an interaction better than the purely socially-motivated rule-based method.

We perform all the experiments following the three-fold cross-validation scheme introduced Section 2.3.2 for each day in the dataset. All methods are evaluated with the same cross-validation splits. For the generalization evaluations in Section 2.5.2, we use data from different days and datasets, instead of the day-wise cross-validation splits to assess the transferability of the model. For all results in Table 2 - Table 8, we report the averaged validation results. For the proposed model, variants with different inputs are all re-trained.

2.5.1 MODEL COMPARISON

In Table 2.2, we show that both the frame-based and sequence-based methods perform better than the rule-based method. While the rule-based method explicitly encapsulates the social dynamics through the speaking status, the rest of the models are able to learn the social rules implicitly without the knowledge of a social prior. We observe that the sequence-based outperforms the frame-based method for Day 1 and Day 3, implying that a temporal context is still beneficial when estimating head orientations. Among all the considered methods, the proposed group-based method achieves the lowest RMSE on average. We perform pairwise t-test to see if the difference in results between the proposed method and the baselines are statistically significant. We find that the group-based method provides statistically lower RMSE than all three baselines for all three days with $p < 0.01$.

We further report the group-size-wise performance of the proposed method in Figure 2.7. Dyads typically have the lowest errors while larger groups have higher errors. This aligns with the intuition and corroborates previous observations [24] that the dynamics within dyads are typically simpler and easier to model compared to that of larger groups.

As the speaking status is a critical component of the rule-based method, we report the listener- and speaker-specific performances in Table 2.3. According to the social rules, the listener’s head orientation is expected to be biased towards the speaker. As the target of attention is clear for the listeners, we anticipate the estimation errors for listeners to be lower than that of the speaker (since the speaker typically divides the attention among several listeners). However, Table 2.3 show that the errors for listeners are consistently higher than those for speakers. Possible reasons are that, in the crowded and noisy mingling scenarios, 1) there are multiple speakers in a group, and 2) people direct their attention

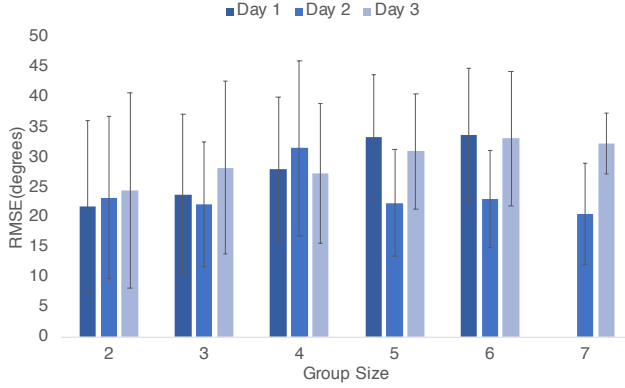


Figure 2.7: Group-size-wise performance of the proposed group-based method in head orientation estimation. Note that there are no groups with seven members in the Day 1.

Table 2.3: Head orientation estimation performance breakdown of the rule-based method on listeners and speakers separately based on results in Table 2.

	Mean RMSE in θ_h (std. dev.) [$^{\circ}$]		
	Day 1	Day 2	Day 3
Listener	51.0 (33.5)	46.9 (30.1)	56.0 (37.0)
Speaker	40.7 (27.0)	36.8 (23.6)	50.3 (35.1)

elsewhere due to distractions, which complicates the interaction dynamics. Using the proposed method, we observe slightly lower performance for the listeners than for the speakers (Table 2.4), with the difference in performance being small compared to the same for the rule-based method.

2.5.2 GENERALIZATION TO UNSEEN DATA

CROSS-DAY GENERALIZATION WITHIN THE MATCHNMINGLE DATASET

The three days for which the MatchNMingle data are available have different distributions of group sizes (Figure 2.4) as well as qualitative differences that include demographics, personality, acquaintance level, etc., of the participants. To this end, we assess the generalization of the proposed model to unseen data by training it on data from one day and testing on the data from the remaining days. The results are summarized in Table 2.5.

Table 2.4: Head orientation estimation performance breakdown of the proposed group-based method on listeners and speakers separately based on results in Table 2.

	Mean RMSE in θ_h (std. dev.) [$^{\circ}$]		
	Day 1	Day 2	Day 3
Listener	22.0 (17.0)	22.8 (13.9)	25.9 (16.9)
Speaker	19.9 (16.4)	18.4 (14.6)	22.5 (16.7)

Table 2.5: Cross-day generalization within the MatchNMingle dataset. Reported are the head orientation estimation performance of the proposed group-based model when trained and tested on data from different days. We use all data from a specific day as training data, and subsequently evaluate the model using data from the other two days as test data.

	Mean RMSE in θ_h (std. dev.) [°]		
	Test: Day 1	Test: Day 2	Test: Day 3
Train: Day 1	N/A	30.6 (29.5)	24.2 (23.4)
Train: Day 2	17.5 (14.8)	N/A	19.6 (16.7)
Train: Day 3	17.3 (14.3)	18.6 (13.6)	N/A

Overall, the model shows strong generalization to unseen data of contextually similar but different social scenarios.

CROSS-DATASET GENERALIZATION TO SALSA-COCKTAIL-PARTY DATASET

As shown in Table 2.1, the SALSA-Cocktail-Party dataset [45] is the most similar to MatchNMingle in terms of data modality. We preprocess the SALSA data analogously to the MatchNMingle inputs. Subjects’ positions, tri-axial accelerometer signals, body orientations, and head orientations are arranged into sequences based on group memberships. The features are available at the frequency of 1/3 Hz. The sequence length is set to $T = 10$ (same as MatchNMingle) which corresponds to 30 seconds in time. Audio data are only available as summary-level statistics and Mel-frequency cepstral coefficients. Obtaining binary speaking status signal from these signals is a challenge in itself and beyond the scope of this generalization test. Hence, we exclude the speaking status from the feature set. To this end, we train the group-based model on the data from the MatchNMingle dataset (we use Day-2 data as training data as it contains a more even distribution of group sizes compared to other days) excluding speaking status from the inputs. Without any re-training or fine-tuning, the model is evaluated on the SALSA data. We report the RMSE in head orientation estimations to be $22.5 \pm 14.0^\circ$, which implies that the proposed model is capable of generalization to different datasets.

2.5.3 CONTRIBUTION OF DIFFERENT MODALITIES

Given our input modalities (positions, speaking status, body orientations, accelerometer signals), we re-train the group-based model on a non-exhaustive combination of inputs relevant to this paper and assess their individual contributions (Table 2.6). Body orientation contributes the most to the performance, which corroborates previous observations in head orientation estimation in pedestrian [117] and social settings like poster session [28]. Predictions based on positional information only are worse than those based on body orientations only. However, combining body orientation and positional information as well as integration of speaking status and accelerometer signals successively improve the performance (the rightmost column shows results from the full set of modalities (features)).

2.5.4 REGRESSION VS. CLASSIFICATION

This paper focuses on estimating head orientation in a regression setting. In this section, we investigate the model performance by framing head orientation estimation as an 8-class

Table 2.6: Head orientation estimation performance of the proposed group-based model with some of the relevant combinations of the modalities (features). Abbreviations BO, pos, SS, and acc stand for body orientation, position, speaking status, and accelerometer signals, respectively. Each column lists the features used in re-training the model.

	Mean RMSE in θ_h (std. dev.) [$^\circ$]					
	pos	pos+SS	BO	BO+pos+SS	BO+pos+acc	BO+pos+SS+acc
Day 1	37.2 (22.7)	35.9 (18.2)	25.1 (13.5)	23.7 (13.8)	23.3 (14.2)	22.7 (14.0)
Day 2	42.1 (24.7)	38.5 (16.2)	26.6 (16.8)	23.6 (12.1)	23.7 (11.9)	22.0 (12.6)
Day 3	40.9 (28.7)	36.9 (22.2)	26.0 (14.4)	26.2 (14.6)	27.4 (15.7)	25.0 (15.9)

Table 2.7: Head orientation estimation performance of the group-based model in a regression vs. 8-class classification setting. Instead of using continuous labels, we use a discretized version of the labels based on 8 classes to train the model and showcase the effect of regression vs. classification.

	Mean RMSE in θ_h (std. dev.) [$^\circ$]	
	Regression	8-class classification
Day 1	22.7 (14.0)	24.5 (15.7)
Day 2	22.0 (12.6)	23.4 (14.0)
Day 3	25.0 (15.9)	27.6 (19.9)

classification task which is the more common approach [28, 50, 88, 118]. To simulate discrete orientation class-labels, we categorize the continuous annotations of head orientations in the MatchNMingle dataset into 8 classes of size equal to 45° each. We train the group-based model using the centers of the bins as labels. During evaluation, we adjust the predicted labels to the center of the corresponding bins and compute the RMSE with respect to the undiscretized ground-truth head orientations. We report our findings in Table 2.7, where we show that estimating head orientation in a regression setting is more accurate over a 8-class classification setting. We also highlight that it is indeed feasible to obtain good continuous regression results from discrete training labels using our method. This is especially promising for application to other datasets, as most of the publicly available ones only have discretized head orientation labels.

2.5.5 USING BODY ORIENTATIONS

We assess the sensitivity of results with respect to manually labeled body orientation inputs. To this end, we replace the body orientations with an approximation based on the interaction space of the group using member positions only. For 2-person groups, the body orientation of each person is approximated by the positional orientation from the respective position towards the mean of the positions. For groups of larger sizes, a circle is fitted geometrically using the Kasa algorithm [119] and the center of the fitted circle is interpreted as the group center. The body orientation of each member is approximated by the positional orientation from the member position to the found group center. This approximation partially relaxes the requirements of body orientation inputs originating from manual annotations, automated vision-based methods, or other wearable-sensing capabilities.

Table 2.8: Head orientation estimation performance of the group-based model with ground-truth (GT) and approximated body orientations. Instead of using the GT body orientations as features, we re-train the model and assess the model performance with the use of an approximated version of the body orientations, based on positions.

	Mean RMSE in θ_h (std. dev.) [$^\circ$]	
	Body orientation (GT)	Body orientation (approximate)
Day 1	22.7 (14.0)	35.6 (22.5)
Day 2	22.0 (12.6)	34.9 (18.3)
Day 3	25.0 (15.9)	36.0 (25.2)

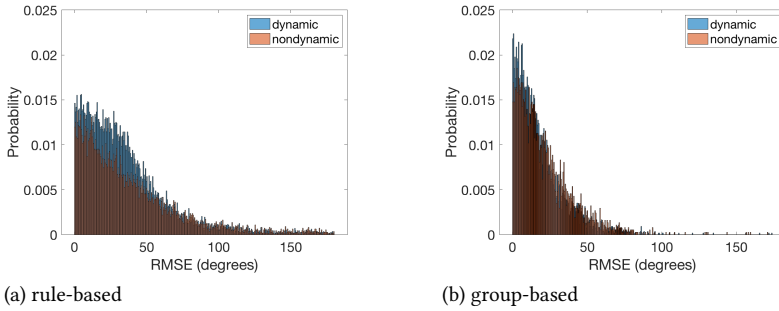


Figure 2.8: RMSE distribution of Day 1 validation results using the rule-based and the proposed group-based approach.

2.5.6 SPEECH DYNAMICS vs. HEAD ORIENTATION ESTIMATION

Including sequences of speech status as the part of the inputs is motivated by how speech plays an important role in group interactions. While positions and orientations information capture the proxemics context, speech activity is related to interaction dynamics.

We propose a scheme to categorize samples as dynamic or nondynamic depending on speaking status switches to further showcase the efficacy of our proposed method. For a given time t , a fixed interval of window w is defined such that the speaking status sequence $[t - w, t]$ is considered. For the speaking status at t , the nearest switch in speaking activity in this time interval is recorded. If the nearest switch is less than w , the sample is considered dynamic and else, it is deemed nondynamic. For this paper, we consider w as 10 seconds, representing the sequence length derived from turntaking duration. An illustration of the performance of dynamic and nondynamic samples from the validation results of Day 1 data is shown in Figure 2.8. Figure 2.8(a) shows that the head orientation estimations obtained from the rule-based method, which is purely driven by speech activities along with position information and prior knowledge from social science. Dynamic samples have larger errors compared to nondynamic samples. Figure 2.8(b) shows the error distribution obtained using the proposed inputs and model. The discrepancy of performance in dynamic and nondynamic samples is greatly reduced, showing the efficacy of our approach in handling dynamic scenarios.

2.6 DISCUSSION AND FUTURE WORK

2.6.1 SOURCE OF INPUTS

A remaining question regarding the efficacy of our proposed method is that of practicality, since it is built on a majority of ground truth inputs. We explain how these inputs (speaking status, body orientation, relative positions, and accelerometer data) may be acquired using existing technology and/or automated methods (except accelerometer data which cannot be replaced by manual annotations) for an end-to-end solution along with some of the associated challenges.

- **Speaking status:** A number of wearable badges measure the speaking status directly [63, 120]. Gedik and Hung [121] have shown that speaking status can also be obtained from a tri-axial accelerometer alone; they reported an Area-Under-Curve (AUC) score of 68% (current state-of-the-art performance via this modality).
- **Body orientations:** Wearable light tags introduced by Montanari et al. [98] allow for direct measurement of body orientation, but are also limited to low sampling frequencies.
- **Positions:** Previous works have demonstrated the use of mobile devices [122], bluetooth beacons [123], radio frequency identification tags [124], etc. to estimate the subject positions. It is still an open research problem to estimate subject positions in dense and crowded settings like the ones we are interested in.
- **F-formations:** Gedik and Hung [24] have demonstrated the use of tri-axial accelerometers to identify F-formation group membership. However, there is room for improvement as they only consider the group dynamics but not the inherent proxemics definition of F-formations.

A seamless integration of different sensors recording multiple modalities in crowded and unstructured mingling scenarios is challenging and would require a custom solutions which address issues such as synchronization and unification of signals across different platforms. While cameras are a possible and easy alternative, occlusions are unavoidable due to constraints arising from camera placements and ceiling height of the room. Additionally, cameras-based solutions are not possible in, e.g., outdoor events or low-light scenes. We argue that a purely sensor-based approach for head orientation estimation has the potential to be extended to a wide range of settings. This paper serves as a step towards that direction.

2.6.2 IMPORTANCE OF PRIOR KNOWLEDGE

While the group-based method (based on a deep-learning approach) gives the best performance, a simple rule-based method (Section 2.3.3) serves as a decent starting point (with approximately 48° error on average). The rule-based method is solely based on the knowledge of the group memberships, positions, and speaking status of the subjects, and does not require large amount of training data (unlike the deep-learning approach). Further development of rule-based heuristics and hand-crafted models (e.g., dynamic Bayesian networks [19, 21, 38, 125, 126]) which leverage the prior knowledge are worth considering if data resources are scarce and if more model interpretability is preferred.

2.6.3 MULTIMODAL FEATURES AND FUSION METHODS

Our group-based model can be adjusted to incorporate more or fewer modalities. Additional relevant modalities such as – gestures, audio, facial expressions, etc., may be introduced depending on availability. An early fusion approach such as the one proposed here may not be directly applicable since the representation of each modality can differ vastly. Early fusion requires the features from multiple modalities to be highly engineered and preprocessed such that they are synchronized and aligned with each other [127]. These problems may be solved by using a late fusion approach. In contrast to early fusion where only one model is trained, different models can be trained for different modalities and all the unimodal representations or decisions are later fused.

2.6.4 GROUP-SIZE-AGNOSTIC vs. GROUP-SIZE-SPECIFIC MODELS

A highlight of the proposed group-based model is that it can be applied to head orientation estimations for groups of different sizes (which are prevalent in real-life social scenarios). However, previous works [24] suggest that the dynamics can be different between small and large groups. While our approach is more general, building group-size-specific models could lend more focused insights into the group interactions.

2.7 CONCLUSION

In this paper, we have proposed an LSTM-based model for understanding the dynamics of joint head motion and behavior in human social interactions, particularly in unstructured and in-the-wild mingling scenarios. Leveraging the implicit coupling between the behavior of the group members, the model jointly predicts the head orientation of all the members solely based on their (temporally evolving) proxemics, conversation dynamics, and body movements. The group context is captured by pooling the hidden states of the group members at each step during the LSTM unrolling. The specific choice of inputs serves as a departure from utilizing visual data (which are limiting due to occlusions and poor lighting) and is a step toward a purely sensor-based and non-intrusive approach for head orientation estimation in crowded and in-the-wild settings. We tested our approach on the MatchNMingle dataset which is based on crowded mingling in casual networking events. Our proposed method outperforms a rule-based method (hand-crafted based on the knowledge of social manners) and deep-learning baseline methods that do not explicitly employ the temporal and group context together. The model demonstrated strong generalization to unseen data across different days of the same event, as well as a completely different dataset (SALSA-Cocktail-Party) without any re-training or fine-tuning. We also showed that the model is applicable to groups of different sizes. Our sensitivity analyses assessing the inputs of speaking status and tri-axial acceleration, in addition to body orientations, showed that these modalities contribute positively towards model performance. We showed that formulating head orientation estimation in a regression setting not only agrees more with the continuous nature of angular data, but is also more advantageous over the more conventional classification setting. We shed some light on possible future improvements of this model, particularly in the direction of using automated inputs and further fusing prior social science knowledge and multimodal signals to better capture the interaction context which affects head orientation estimation.

3

3

CONVERSATION GROUP DETECTION WITH SPATIO-TEMPORAL CONTEXT

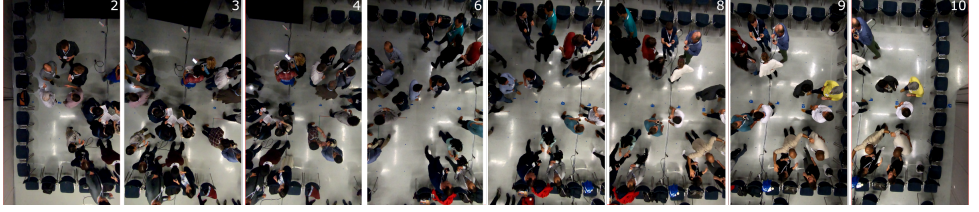


Figure 3.1: Snapshot of a typical social interaction area from the Conflab [128] dataset. Faces blurred for privacy.

3

3.1 INTRODUCTION

The automatic detection of conversation groups is an interesting problem for applications ranging from social surveillance [6, 17, 129] to social robotics [130–132]. In social settings, such as cocktail parties or professional networking events, the floor for interactions at the venue consists of multiple conversation groups that dynamically adapt to the ebb and flow of the underlying human behaviors that determine the social interactions. Characterizing the interpersonal relationships (the level of spontaneous affinity between people) that foster individuals to freely congregate to form a focused encounter and exchange information could help us understand more about interaction experience and quality [133, 134]. However, due to the complexity and subtleties in human social dynamics in changing environments which are also context-specific, automatically detecting conversation groups in social scenes is an ongoing and challenging research topic.

For this paper, we focus on identifying conversation groups (more specifically, free-standing conversation groups (FCGs) [6]) in a social scene where individuals physically come together to interact with each other. In these settings they organize themselves into groups that define the physical partitions of who is interacting with whom. Figure 3.1 shows a representative scene for the types of social scenarios that we are interested in from an overhead viewpoint. People’s use of physical space is known as *proxemics* [135]. The conversation groups can be of varying duration, size, and spatial arrangement. In practice, these conversation groups have often been conceptually formalized as *facing-formation* (F-formations) [6, 17, 28]. As introduced and defined by Kendon [23], an F-formation is formed when two or more people arrange themselves onto a convex envelope to enclose an overlap of their interaction transactional segments (i.e., space in front of them where sight and hearing are most effective [136]). The interactants have equal and exclusive access to this overlap (i.e., *o-space*).

The challenges in formulating a social concept such as F-formations into a computational task for automated machine learning methods are two-fold: (i) representation of the scene and identifying the appropriate behavior cues to capture the underlying social dynamics, and (ii) the potential fuzziness that exists in the ground truth of group membership due to the fact that interpersonal relationships are not exactly binary in reality.

A social scene such as Figure 3.1 may be represented by an interaction graph with nodes representing the individuals, and edges representing the relationship between two individuals. Conversation groups can be deduced from the information in this interaction graph, where individuals in the same conversation have greater edge weights (affinity) with one another, and vice versa. Each individuals have behavioral cues such as positions,

head/body orientations, etc. that are shaped by the surroundings, and also change over time. Indeed, proximity with directional information could already be indicative of group membership if one simply considers 'who is standing with whom'. However, in social scenes such as Figure 3.1, factors such as crowdedness and furniture layout define the spatial context that affects individual cues that determines conversation group membership. The temporal context in the behavioral cues plays a role in the interactions which are dynamic in nature. One type of social dynamics that could be captured by movement cues (e.g., change in positions and orientations) is synchrony and mimicry patterns which are known to be important driver for affiliation and interpersonal rapport [137–139]. In terms of conversation dynamics, since multiple conversation floors are possible within one F-formation [81], head orientations that change during turn-taking or other types of conversation dynamics are especially relevant to conversation schisms. The schisming phenomenon may lead to two or more distinct conversation groups [140] and hence possibly form new F-formations. Therefore, in this paper, we argue that even though F-formation in its original definition is a static concept, it is important to take into account the temporal context of the behavioral cues influencing the groups.

The second fold of the challenge is the potential problem in how ground truths of group membership are defined from pairwise affinities. Past approaches of automatic detection of conversation groups have identified affinity and group memberships as binary and operationalized the task based on this design choice, albeit Zhang and Hung have investigated the subjectivity of annotating groups [42]. Although the assumption of the binary group membership which existing methods and ours hinge on, is valid to the extent of how they are reflected in the ground truth annotations, interpersonal relationships are not binary in reality. The temporally evolving affinity between two individuals, does not change from zero to one, or one to zero, instantaneously [141–143]. Social interactions have a rite of passage, from greeting to leaving [144]. We aim to understand how the affinity scores change in time and encapsulate the changing dynamics of the behavioral cues, and how they affect group detection, which are not apparent in hard assignments of group memberships. Our paper takes a step towards this direction which has not been the focus in previous works.

Following prior works [59, 60], our approach to conversation group detection consists of two stages: (1) we first estimate continuous pairwise affinities between all individuals in a social scene, and (2) we use an existing framework to cluster the individuals by leveraging a graph clustering based on Dominant Set (DS) to identify conversation groups [17]. In order to account for the temporal context in the behavioral cues, we introduce a deep learning based Long Short-Term Memory (LSTM) network to predict pairwise affinity scores that determine the F-formation membership (annotated as ground truth of conversation groups in existing datasets). The inputs to our network are temporally aligned sequential inputs (including positions, and head and/or body orientation); the output of our network is the pairwise affinity value between one individual and *all* other members in the social interaction scene.

In contrast to existing works that output an affinity score of all pairs of individuals independently which scales quadratically with the number of people in the scene [59, 60], our design for person-wise output scales linearly and follows more intuitively from an egocentric application point of view (i.e. social robots). DS clustering is applied to the

affinity matrix corresponding to each scene to detect the conversation groups. The clique formulation in DS is exploited to identify clusters in the interaction graph, which refines group detection since the affinity matrix (from estimated pairwise affinities) may not directly provide self-consistent and symmetric binary group memberships [17, 60, 145].

Additionally, as opposed to previous approaches that only use the intermediate affinity scores as inputs for clustering for refined conversation group identification, we show the possibility in using the estimated pairwise affinity values from the past to forecast future affinity values, which could serve as an underpinning for understanding how conversation groups evolve. Even without formulating the problem explicitly as a future forecasting task, our model is able to anticipate affinity values due to the temporal continuity. Our main contributions are as follows:

- We propose a novel LSTM-based affinity score model to approximate the likelihood of two people interacting in the same conversation group. The model includes a pooling module to account for the spatial context of social interactions, inspired by what [59, 60] captured in their models. Using the proposed model that leverages (temporal) sequential input features in addition to a pooling module, we simultaneously account for the spatio-temporal context that affects pairwise affinities that determine the conversation groups.
- We provide an analysis over the predicted affinity scores in conversation group detection, characterized by Area under (receiver-operating) Curve (AUC) scores, followed by qualitative examples showing the continuity of affinity scores. We also show a comparison of affinity score processing (asymmetric vs. symmetric) for application in DS and the group detection performance with respect to scene dynamics.
- We demonstrate the usability of the predicted affinity values via a novel forecasting framework for affinity score prediction based on Gaussian Process Regression (GPR). The framework also provides inferential uncertainty quantification over the predictions of future conversation groups.

3.2 RELATED WORKS

Conversation group detection in situated interactions has been tackled by a variety of approaches stemming from different communities (computer vision [6, 50, 129], human-computer interaction [145–149], etc.). This section discusses the representative works in this area. Conversation groups and the more formalized F-formation representation are analogous in terms of group detection in interaction scenes in past works.

Many previous works, especially from the vision community, use features such as location and head/body orientations for the task of group detection [6, 17, 28, 50, 53]. These quantities could be obtained automatically from vision data using multi-camera surveillance setups (typically elevated side-views).

Using these features, some methods for F-formation detection have been focused on optimization-based approaches to mathematically model the physical space. More concretely, a number of works [6, 150, 151] hypothesised that the o-space can be generated from a noisy representation of the instantaneous view frustum obtained from the head

pose. Heat maps generated from samples projected from each individual's view frustum were then used to identify o-spaces. Members of the F-formation were then re-identified as belonging to a particular o-space based a pre-defined metric of closeness.

Another class of methods have formulated social scenes as an edge-weighted graph where each individual represents a node and the edge represents the pair-wise connection between individuals. These methods take the part of the F-formation definition related to equal mutual attention to synonymous with maximal cliques in edge weighted graphs. In early works, the pair-wise relationships were modeled using feature engineering based on location and orientation information [17, 42, 53, 152]. Aggregating these estimated pairwise affinity values, the affinity matrix serves as inputs to graph clustering based on Dominant Set using game-theoretic approaches [17, 53] to iteratively partition nodes to extract conversation groups. While the o-space is not explicitly modelled with these approaches, the maximal clique formulation implicitly models the o-space whilst also explicitly binding individuals to a specific group as part of the Dominant Set identification process. However, the representation of pairwise affinity, particularly when only location and orientation is used, forces a circular shape assumption to the F-formation that does not always happen in practice [42, 152]

To address this problem and enable more flexibility in modeling pairwise relationships given the surroundings, deep learning based approaches have been proposed. DANTE learns the affinity values by explicitly modeling dyadic and context interaction [60] by using relative positions and head/body orientations after preprocessing as inputs to the model. More recently, Thompson et al. [59] proposed a graph neural network (GNN) based approach that leverages the more general message-passing mechanism during training to predict affinities using raw signal data including absolute positions, accelerometer readings, and image. Similar to the preceding works [17, 42, 53], both of these deep learning based approaches also apply the learned affinities values inputs to DS graph clustering. Schmuck and Celiktutan [153] also proposed a GNN-based approach to predict interpersonal links, but as opposed to using DS graph clustering, a greedy agreement algorithm was applied to identify groups [145].

Departing from using visually obtained features such as locations and orientations of individuals, some works have taken advantage of features from other modalities that have shown to be helpful when estimating pairwise affinities. For example, [59] take advantage of a combination of motion based features and visually obtained features. Gedik and Hung have shown that it is possible to estimate groups purely based on motion features as phenomena such as body movement synchrony in interactions are indicative of pairwise relationships [24]. However, in communication with the authors, the predicted affinities need to be significantly improved for before DS clustering would yield reasonable performance. This highlights that the nature of the problem lies between the modelling social dynamics and proxemics (i.e. positions and orientations). Approaches proposed by the ubiquitous and pervasive computing communities have relied on Bluetooth Smart (BLE) to measure proximity values in terms of Received Signal Strength Indicator (RSSI) values which capture distance and orientation information to some extent, represented by [154]. In the case of [155], data from motion sensors (accelerometer and gyroscope) were also incorporated with proximity features for multimodal detection of groups using smart phones. Other custom sensors have been developed to measure proximity, relative

orientation, motion, and/or a combination thereof (e.g., light tags [156], Rhythm badges [157] and the Midge [158]). These data also capture the useful information, such as direct measurement of closeness forming hypotheses of interactions already and the measurement of nuanced body motion, in determining conversation groups in social interactions, and methods developed based on these have the potential to scale more easily.

The proposed method differs from early works [17, 42, 53, 152] in F-formation detection as it is a deep learning method where pairwise affinities are learned. The main limitation is that the early works are based on feature engineering, which is less flexible than deep learning models. Compared to existing deep learning methods [59, 60, 153], the proposed method simultaneously models the spatio-temporal context in F-formation detection. We aim to capture the underlying evolving social dynamics by modeling the temporal dynamics in the input features, which has not been investigated before. Our method outperforms the most relevant deep learning baseline [60] for the ConfLab dataset, especially in scenes where there is high dynamicity.

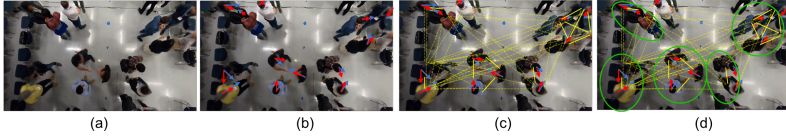


Figure 3.2: Visualization of the derivation of the interaction graph: (a): the given scene from an overhead camera; (b): each subject has person-wise features, such as positions, orientations, etc.; (c): the edge weights (affinity scores) between each pair of subjects are predicted via the proposed neural network model; and (d): using the DS [17], conversation groups are extracted as sets of subjects.

3.3 APPROACH

The overview of the approach to conversation group detection is illustrated in Figure 3.2. Figure 3.2(a) represents an example image from an interaction scene. The individual attributes such as positions, head and body orientations encode spatial information of an individual with respect to the scene (labeled in Figure 3.2(b)).

Module (b) represents the core of our contribution, which is a novel deep learning neural network for pairwise affinity estimation, based on a joint Long Short-Term Memory (LSTM) network that simultaneously accounts for the temporal context of the input signals and spatial context in the scene with spatially-motivated context pooling. In (c), the pairwise affinities are combined to a affinity matrix, and following previous approaches, Dominant Set was used to extract groups by iteratively identifying maximal cliques in edge-weighted graphs (module (d)). Our contribution focuses on the neural network architecture for affinity prediction, and assumes that the inputs are acquired and preprocessed upstream. We use the Dominant Set clustering method on graphs downstream of affinity prediction because it is state-of-the-art method for this use-case.

The details of the neural network architecture is described in Sec. 3.3.1, and the details of the dominant set method is described in in Sec. 3.3.2.

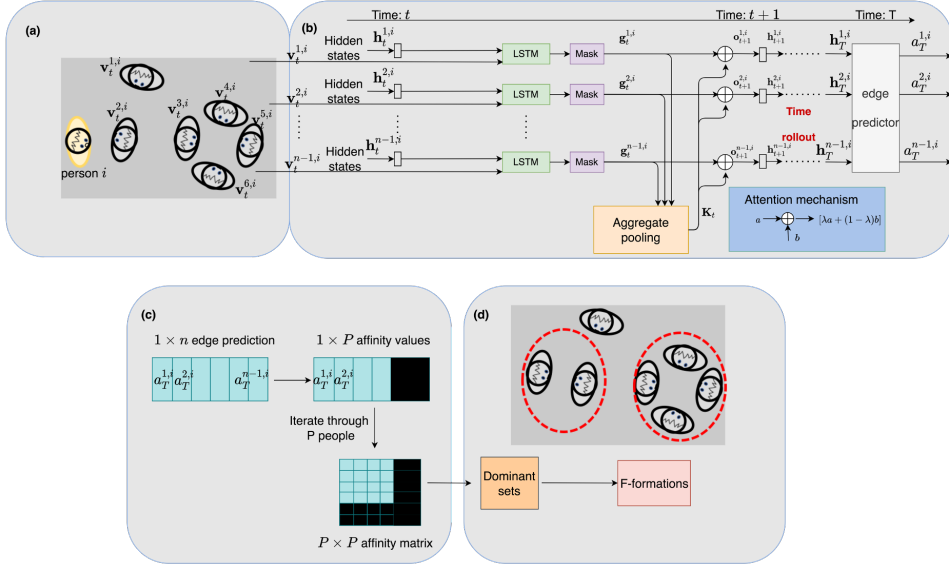


Figure 3.3: Visualization of the methodology. (a): a scene representation of individuals and preprocessed features with respect to person i ; (b): a graphical representation of the time rollout from t to T with the aggregate pooling layer along with the attention mechanism; (c) the filtering and aggregation of person-wise affinity output at $t = T$ into an affinity matrix; (d): application of DS clustering for group identification from the affinity matrix.

3.3.1 AFFINITY PREDICTION

For a given social interaction scene S , let P_t represent the number of individuals in the scene at time step $t \in \{1, \dots, T\}$ where T is the sequence length, and n represent the maximum number of individuals present at all scenes of concern, i.e., $P_t \leq n$. Let $v_t^{j,i} \in \mathbb{R}^N$ denote the N -dimensional feature vector for the i^{th} member of the scene with respect to the j^{th} member (with $1 \leq i, j \leq P_t$ and $i \neq j$) at the sequence step t . The feature vector is a concatenation of features based on the following:

- head and/or body orientations of member i ,
- position, and head and/or body orientations of all members j relative to the member i ,
- indicator mask $I_t^j = \{1, 0\}$ – denoting if member j is present in the scene at time t (assumed to be known a priori),

the details of which are discussed in Section 3.4.4.

Module (b) in Figure 3.3 demonstrates one recurrent step of the proposed model from time t to $t+1$. Let $h_t^{j,i} \in \mathbb{R}^H$ denote the hidden representations associated with member i at sequence step t , where H is the dimension of hidden states (chosen as a hyperparameter). The hidden states at $t=0$ are initialized as $h_0^{j,i} = \mathbf{0}$. To capture the spatial context defined by all members in the interaction scene, we pool the hidden states of all the members $j \neq i$ as follows. To discount for the persons absent in the scene at time t , the hidden

state are first processed through a masking layer (with element-wise multiplication) to obtain intermediate masked hidden states representation $\mathbf{g}_t^{j,i} = I_t^j \mathbf{h}_t^{j,i}$. Note that the masked representation reflects the presence of members j in the scene, which may be dynamically changing between different time steps. All masked representations $\mathbf{g}_t^{j,i}$ are processed through an *aggregate pooling* layer to obtain a scene level representation \mathcal{K}_t^i given by

$$\mathcal{K}_t^i = \frac{\sum_{j \neq i} \mathbf{g}_t^{j,i}}{\sum_{j \neq i} I_t^j}. \quad (3.1)$$

The context-pooled representation \mathcal{K}_t^i is then combined with each of the individual hidden states $\mathbf{g}_t^{j,i}$ through an attention mechanism to obtain $\mathbf{o}_t^{j,i}$,

$$\mathbf{o}_t^{j,i} = \lambda \mathcal{K}_t^i + (1 - \lambda) \mathbf{g}_t^{j,i}, \quad (3.2)$$

where $\lambda \in [0, 1]$ is a trainable parameter that adjusts the contributions from pairwise interaction and surrounding representations.

For each member i , the respective hidden state $\mathbf{h}_t^{j,i}$ as well as the concatenation of the feature $\mathbf{v}_{t+1}^{j,i}$ and the processed context representation $\mathbf{o}_t^{j,i}$ are passed through an LSTM cell \mathcal{L}_τ (parameterized by τ) to obtain $\mathbf{h}_{t+1}^{j,i}$ (i.e., the hidden states for the subsequent time step),

$$\mathbf{h}_{t+1}^{j,i} = \mathcal{L}_\tau([\mathbf{v}_{t+1}^{j,i}, \mathbf{o}_t^{j,i}], \mathbf{h}_t^{j,i}). \quad (3.3)$$

The LSTM operation \mathcal{L}_τ is described by the following series of transformations

$$\begin{aligned} \mathbf{f}_{t+1}^{j,i} &= \sigma(\mathcal{W}_{\xi_f}([\mathbf{v}_{t+1}^{j,i}, \mathbf{o}_t^{j,i}, \mathbf{h}_t^{j,i}])) && \text{(forget gate's activation vector)} \\ \mathbf{i}_{t+1}^{j,i} &= \sigma(\mathcal{W}_{\xi_i}([\mathbf{v}_{t+1}^{j,i}, \mathbf{o}_t^{j,i}, \mathbf{h}_t^{j,i}])) && \text{(input gate's activation vector)} \\ \mathbf{o}_{t+1}^{j,i} &= \sigma(\mathcal{W}_{\xi_o}([\mathbf{v}_{t+1}^{j,i}, \mathbf{o}_t^{j,i}, \mathbf{h}_t^{j,i}])) && \text{(output gate's activation vector)} \\ \tilde{\mathbf{c}}_{t+1}^{j,i} &= \tanh(\mathcal{W}_{\xi_c}([\mathbf{v}_{t+1}^{j,i}, \mathbf{o}_t^{j,i}, \mathbf{h}_t^{j,i}])) && \text{(cell input activation vector)} \\ \mathbf{c}_{t+1}^{j,i} &= \mathbf{f}_{t+1}^{j,i} \odot \mathbf{c}_t^{j,i} + \mathbf{i}_{t+1}^{j,i} \odot \tilde{\mathbf{c}}_{t+1}^{j,i} && \text{(cell state vector)} \\ \mathbf{h}_{t+1}^{j,i} &= \mathbf{o}_{t+1}^{j,i} \odot \tanh(\mathbf{c}_{t+1}^{j,i}) && \text{(output hidden state vector),} \end{aligned} \quad (3.4)$$

where σ is sigmoid activation, \odot denotes the Hadamard product, and $\mathbf{c}_t^{j,i}$ and $\mathbf{c}_{t+1}^{j,i}$ denote the cell state at t and $(t+1)$, respectively. $\mathcal{W}_{(\cdot)}$ denotes a linear layer with parameters indicated in the subscript. The trainable parameters are contained in the set $\tau = \{\xi_f, \xi_i, \xi_o, \xi_c\}$. Importantly, the LSTM parameters τ are shared among all the members in the scene.

After the time roll out in each LSTM time step until $t = T$, the hidden states $\mathbf{h}_T^{j,i}$ are passed through a linear layer parameterized by set of weights \mathbf{W}_m and biases \mathbf{b}_m to obtain the final pairwise edge predictions with respect to member i . Subsequently, they are passed through a sigmoid activation (denoted by σ) to obtain pairwise affinity $a^{j,i} \in [0, 1]$ as

$$a^{j,i} = \sigma(\mathbf{W}_m \mathbf{h}_T^{j,i} + \mathbf{b}_m), \quad (3.5)$$

where the values of 0 and 1 denote no and perfect pairwise affinity, respectively. The output of the model $a^{j,i}$ is strategically designed to be continuous, which lends naturally to

a probability interpretation of pairwise interaction. We further motivate this choice, the performance evaluation, and the connection to downstream tasks such as conversation group forecasting in Section 3.4.3.

To train the model, we use the mean squared error loss function given by

$$\ell = \sum_S \sum_{j \neq i} (a^{j,i} - a_{GT}^{j,i})^2, \quad (3.6)$$

where $a_{GT}^{j,i} \in \{0, 1\}$ represents the ground truth affinity value between member j and member i .

3

3.3.2 DOMINANT SET CLUSTERING

As shown in Figure 3.3 (c), for each member i in the scene, the affinity prediction model predicts $n - 1$ pairwise affinity values with respect to all other members j for all time steps irrespective of whether they are visible at that moment or not. To evaluate the group identification performance, we use the output at the last time step at $t = T$. After filtering with the indicator masks, a $P \times P$ affinity matrix for the scene in question is obtained, where P is the actual number of subjects at a particular scene.

After the predicted affinities are arranged into an affinity matrix, following prior approaches, [17, 53, 59, 60], the F-formations are extracted using Dominant Sets (DS) clustering. The resulting clusters representing F-formations could be of any size. The stopping criterion of the optimization formulation is either when the relative mutual affinity of internal nodes and external nodes of a dominant set do not satisfy the constraints, or when the mutual affinity of a group is lower than a chosen threshold. The second part of the stopping criterion enables improvement to detect singletons in the scene as it accounts for the global context (i.e., when there are only few people left after maximal clique extraction iterations, it is not likely that they are in the same group). We follow the implementation of F-formation clustering of [60]. For the theoretical background and more detailed reference to the application of Dominant Set Clustering for F-formation detection, please refer to [17, 159].

3.4 EXPERIMENTAL SETUP

3.4.1 BASELINE METHODS

The baseline methods considered in this work are GTCG [53], GCFF [6], and DANTE [60]. GTCG and GCFF are both non deep learning based methods which rely on engineered position and orientation based features. GTCG models pairwise affinity values using distance between distributions over the plausible regions determined by the visual frustums, followed by a refined game theoretic approach for group extraction based on [17] and [22]. GCFF models the probability of individuals belonging to o-space centers (i.e. center of conversation groups), and uses a graph-cut approach in conjunction with constraint based on direct access to extract the groups.

DANTE proposes a deep learning based approach to model pairwise affinities using positions and orientations, and utilizes the Dominant Set clustering for extracting conversation groups. During training of the deep learning model, DANTE uses data augmentation strategy. The recently proposed graph neural network based approach takes advantage

of image-based features and a rich collection of social action semantic labels, in addition to proxemics and body motion based features, for conversation group detection [59]. We omit comparison against this recent approach for the scope of the paper, since the focus of our paper is on modeling the social dynamics in conversation scene using temporal information rather than a thorough investigation of using different input modalities.

3.4.2 DATASETS

To align with the closest state-of-the-art approaches, we compare our method on the same representative datasets including Cocktail Party [46] and SALSA [45]. We also report benchmark results on the recently released Conflab dataset [128], capturing professional networking social interactions in-the-wild.

With conversation groups annotated at 1Hz and behavioral cues sampled and annotated at 60Hz then summarized to 1Hz, the Conflab dataset is apt to investigate our research question which is leveraging temporal continuity in behavior cues and pairwise relationships in estimating affinity scores. In comparison, the conversation groups and behavior cues in Cocktail Party and SALSA are annotated at 1/5 Hz and 1/3Hz, respectively. We hypothesize that the temporal continuity in the signals and the ground truth of these two datasets can be decimated due to this sampling and annotation frequency.

Most of the existing datasets were collected to serve F-formation detection using visual information, i.e. using an elevated side-view. Bounding boxes and head/body orientations are acquired either through automated methods or manual annotations. For datasets that have a top-down view, positions and orientations are acquired through manual annotations because automated methods result in error prone inputs to subsequent models [128]. The overview of the datasets used is as follows:

- Cocktail Party* [46]: contains 30 minute recordings of six people interacting with one another, captured by four elevated side view cameras in the corners of the space. Positions and head orientations of the subjects are obtained automatically using a particle filter-based body tracking method. The conversation groups were annotated at 1/5 Hz.
- SALSA* [45]: contains 60 minute recordings of 18 people interacting with one another, captured by four elevated side view cameras in the corners of the space. Positions, head and body orientations, conversation groups of the subjects are annotated manually at 1/3 Hz. This dataset contains wearable sensor data captured by the Sociometric badges.
- Conflab* [128]: contains 15 minute recordings of 49 people interacting with one another, captured by 5 (non-overlapping) overhead cameras. Positions, head and body orientations of the subjects are annotated manually at 60Hz. Note that even though locations and orientations can be acquired automatically, some previous works have pointed out that the automatic methods produce erroneous results, especially in orientation estimation [46]. To avoid confounding sources of errors in these behavioral cues which are quite nuanced as we motivated in Section 3.1, we follow other works that have relied on provided ground truth data as inputs assuming that these will be provided upstream during application [160]. Conversation groups are annotated manually at 1Hz.

3.4.3 EVALUATION METRICS

We evaluate on both stages of our model: (1) pairwise affinity estimation and (2) group detection. For pairwise affinity prediction, the neural network is trained with binary ground

truth, but the predicted affinities are continuous values between 0 and 1, which enables us to do further analysis using these affinity scores. Given the dynamic nature of the proposed model, we introduce an additional evaluation compared to the existing state-of-the-art methods. Existing methods (e.g., [59, 60]) omit assessing the learned affinities only and use them directly for F-formation detection via DS clustering. We argue that there may be nuances in the learned pairwise continuous valued affinities that may anticipate changes in the group membership that may not be apparent from the hard cluster assignment.

The evaluation metric for affinity estimation is the Area Under Curve (AUC) score of the Receiver Operator Curve (ROC). We use AUC due to the high imbalance of the data; there are typically far fewer positive pairwise memberships than negative in the entire scene.

For the second stage of evaluating group detection, we used the standard evaluation metric used in prior work [17, 22, 42, 53, 60] which involves considering an entire group in the ground truth as a single sample. A detected group k is considered to be correctly estimated if $\lceil Thr * \|g_k\| \rceil$ of the members are correctly estimated, where $\|g_k\|$ indicates the cardinality or the size of the ground truth group, and $\lceil x \rceil$ rounds x to the next largest integer. The threshold $Thr \in [0, 1]$ tunes the tolerance of the evaluation to the number of mis-attributed members in a group. It is commonly set to $Thr = \frac{2}{3}$ or $Thr = 1$, representing greater than majority overlap at 67% and complete overlap with the ground truth membership, respectively. A True positive (TP) is therefore any correctly detected group; false negative (FN) is a missed group; and a false positive (FP) is an estimated group that does not exist in the ground truth. The metrics for group detection performance is then computed using F1 measure over the entire image scene which could contain multiple groups.

3.4.4 IMPLEMENTATION DETAILS

In the case of Conflab, the features were extracted to align with the ground truth at 1Hz by averaging all 60 samples before the label. While it is desired to use a higher frequency signal, this preprocessing step should already allow capturing of social dynamics that exist on a second level, such as synchrony and convergence patterns [161, 162], the associated postural sways [163], and some turn-taking dynamics (e.g., turn transitions) [164].

The head and/or body orientation is given by the angular direction of the person's body in $(-\pi, \pi]$. The relative positions of the group members are given by the radial distance (measured in camera or pixel coordinates) and angular orientation in $(-\pi, \pi]$. The circular mean of the body orientations of all the members in the scene are computed as a zero-degree reference for the scene. This addresses the discontinuity in angles as they wrap around $(-\pi, \pi]$. All orientation related features are corrected by the same zero-degree reference. All features are normalized to $[0, 1]$ via min-max scaling. Labels for conversation group membership are annotated manually and the annotation method are described in each dataset respectively.

Similar to the experimental set up in DANTE, due to the small dataset size, all results are obtained by averaging the test splits using 5-fold cross validation. The validation split is selected such that it separates the training set as much as possible from the test data in time. The test data of a fold is used for results whereas hyperparameters are selected based on validation data. The hyperparameters are hidden representation dimension of the LSTM and sequence length of the input sequence. For the Conflab dataset, experimental

results are obtained for all cameras (camera 2, 4, 6, and 8).

Since P_t changes dynamically and the model is trained with a fixed size input using the maximum number of people in all scenes n , we pad the feature vector from P_t to n with dummy values of -1 . As part of the feature vector, the indicator mask variable represents if a member is present at time t such that the aggregate pooling layer does not account for the dummy subjects.

3.5 RESULTS AND DISCUSSION

3.5.1 OVERVIEW

To ensure a fairer comparison with existing methods, we use the same position and head and/or body orientation based feature set of the individuals in the scene for all methods. Table 3.1 shows an overview of the results on baseline methods on the Cocktail Party, SALSA, and Conflab datasets. As we expected, the proposed method outperforms the baseline methods on the Conflab dataset because of the finer temporal granularity in the behavioral cues and group labels. As opposed to DANTE which excels in both Cocktail Party and SALSA dataset, the proposed method may not have leveraged the temporal context when the social dynamics is undersampled.

Table 3.1: F1 scores comparison $Thr = 1$ across different methods. Standard deviation over test samples are shown in parenthesis. *: results reported by [60]; †: results reported by [59]

Method	Cocktail Party	SALSA	Conflab
GTCG	0.29 (-) *	0.44 (-) *	0.40 (0.12)
GCFF	0.64 (-) *	0.41 (-) *	0.31 (0.23)
DANTE	0.58 (0.43) †	0.65 (-) *	0.66 (0.35)
Proposed	0.48 (0.40)	0.46 (0.23)	0.73 (0.31)

To assess whether or not the efficacy of the proposed model for the Conflab dataset is indeed associated to the frequency in conversation group labels, we subsample the Conflab dataset to 1/5 Hz, to match that of the Cocktail Party dataset. The F1 performance at $Thr = 1$ on this subsampled version of the Conflab dataset is 0.58 with standard deviation 0.32. So we see that even with the same setting, there is a decrease in performance due to an undersampling of key dynamic information that is leveraged by our proposed model.

We observe that results of GTFF and GTCG decrease on the Conflab dataset compared to Cocktail Party and SALSA. This may be because the number of people in the scenes of Conflab are dynamically changing, as opposed to the fixed number of people in both Cocktail Party and SALSA (6 and 18 people, respectively), it may be harder to model o-space and determine overlapping transactional segments using a single parameter (stride), as participants' occupancy of floor space changes. DANTE still performs relatively well on Conflab as it also takes into account the spatial context of the surroundings. In addition to modeling the spatial context similar to DANTE, the proposed model relies on the sequential nature of the LSTM-based network to capture inherent temporal dynamics.

With increased performance in affinity estimations (i.e., the model output), the performance in Dominant Set clustering for group extraction also improves. As we argue that the affinity estimations are critical not only because they are inputs to DS clustering, but

also contain valuable information on how pairwise relationships change continuously over time, we include a more detailed analysis of the affinity values and their relationship with the DS clustering step in the next sections.

3.5.2 ANALYSIS OF AFFINITY VALUES

To uncover where the difference in group detection results in Table 3.1 originates from, this section includes an analysis of where the proposed model differs from DANTE in terms of the predicted affinity value for test sets of the Cocktail Party and the Conflab dataset. Table 3.2 shows a comparison of the predicted affinity value results from DANTE and the proposed method using the AUC metric. With its data augmentation strategy and benefiting from the pairwise output setup, the frame-based DANTE strategy works better for the sparsely sampled Cocktail Party dataset. For the Conflab dataset with the higher sampling frequency, the proposed approach takes into consideration the temporal continuity of labelled cues and affinity values with the data-efficient person-wise training and output, resulting in improved AUC scores that led to improved conversation group detection F1 scores. We posit that the temporal granularity could be too coarse in datasets such as the Cocktail Party for the proposed sequential model to be effective.

Table 3.2: AUC results of DANTE and the proposed method for the Cocktail Party and Conflab dataset.

	Cocktail Party	Conflab
AUC (DANTE)	0.92	0.91
AUC (Proposed)	0.83	0.93

Figure 3.4 shows a qualitative example of how the affinity values from the proposed model change temporally as a new conversation group (Subject 2 and 3) forms. We focus on the right side of the interaction floor in the sequence of the scenes shown. The groups provided by the ground truth, predictions from the proposed method, and DANTE are illustrated in the second column. The affinity scores between Subject 1 and 2, and Subjects 2 and 3 from the proposed method are visualized in the third column. The color and value correspondence is shown in the legend. The pairwise affinity scores between Subject 1 and 2 decrease over time, whereas the score between Subject 2 and 3 increase over time.

3.5.3 AFFINITY SCORES IN DOMINANT SET CLUSTERING

As the predicted affinity scores are continuous values, they are not perfect to directly extract group memberships. The pairwise values might differ and this results in discrepancy [60, 145]. Whether to symmetrize and how to symmetrize the predicted pairwise affinity scores is a design choice not thoroughly assessed previously. Options include using the asymmetric raw predicted affinity values, taking the minimum, average, or maximum of the pairwise affinity values. The (a)symmetry could be illustrative of the individual's intention in interacting with the other person, and affect the group clustering performance as this factor may have also affected how the annotators perceived group memberships.

In Table 3.3, we show the sensitivity F1 scores of symmetrizing the affinity matrix using different strategies for the Conflab dataset. The results show that averaging the pairwise affinities leads to improved F1 score in group detection at $Thr = 1$. This implies that while

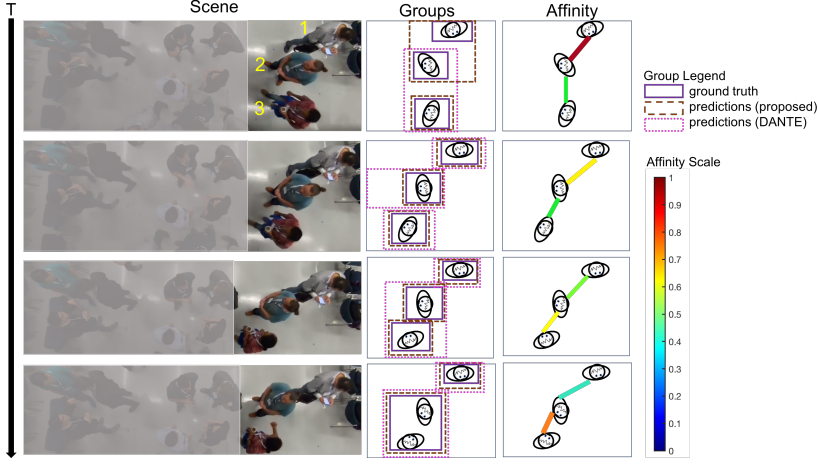


Figure 3.4: Qualitative example of how affinity values change temporally, in relation to a newly formed conversation group.

the asymmetric values are interesting in differentiating people’s likelihood in interacting with each other, a symmetric and averaged representation is better aligned with the binary group membership.

Table 3.3: F1 scores comparison ($Thr = 1$) for the Conflab dataset between different strategies of processing the affinity scores as inputs to DS clustering. Standard deviation over test samples are shown in parenthesis.

	raw	average	minimum	maximum
F1	0.69 (0.33)	0.73 (0.31)	0.72 (0.30)	0.65 (0.35)

Table 3.4: F1 scores comparison ($Thr=1$) for the Conflab dataset with respect to scene dynamics quantified by D . Standard deviation over test samples are shown in parenthesis.

	$D = 1$	$D = 2$	$D = 3$	$D = 4$	$D \geq 5$
DANTE	0.69 (0.36)	0.65 (0.36)	0.78 (0.27)	0.68 (0.32)	0.66 (0.23)
Proposed	0.71 (0.34)	0.69 (0.34)	0.77 (0.31)	0.73 (0.33)	0.77 (0.32)
Delta	0.028 (0.33)	0.037 (0.32)	-0.004 (0.27)	0.039 (0.33)	0.1 (0.35)

3.5.4 PERFORMANCE WITH RESPECT TO SCENE DYNAMICS

We highlight the efficacy of the proposed method when estimating groups especially in scenes that contain more instances of group formations, breaking, and reforming. These events quantify scene dynamics as they imply changes in one or more conversation group reorganization. We define these events based on group presence in the past and future (i.e. a new group is formed if it doesn’t exist before; a group is broken if it doesn’t sustain to the next time step; and a group is reformed when it exists but breaks in the past and now the same members reunite).

We characterize the scenes in the Conflab dataset using this measure of scene dynamics. In Table 3.4, we showcase the performance difference (indicated by Delta) of group detection performance using F1 scores at $Thr = 1$ between DANTE and the proposed method. Delta is calculated by the F1 scores obtained from the proposed method subtracted by that of DANTE. We show the results at different level of scene dynamics denoted as D , where D is the sum of all instances of group formations, breaking, and reformation. We observe

a slight upward trend of the proposed method’s improved performance (i.e., Delta) as scenes become more dynamic (with the exception of $D=3$). When $D \geq 5$ (corresponding to high scene dynamics, whereas $D = 0$ for most of the time where groups are stable), the advantage of the proposed method over DANTE on average is at 0.1. This further shows that the temporal context before a group event takes place may be beneficial in estimating conversation groups.

3.5.5 CONVERSATION GROUP FORECASTING

Using the temporal context of pairwise affinity scores, we further introduce a conversation group forecasting framework. Given a sequence of edge weights a_t^{edge} where $t = 0, 1, \dots, T$, which is the averaged value between each pair of individuals i and j (i.e., $\frac{a_{i,j} + a_{j,i}}{2}$), we predict $a_{t+1}^{\text{edge}}, a_{t+2}^{\text{edge}}, \dots, a_{t+Z}^{\text{edge}}$ where Z is the time forecast horizon. For each a_t^{edge} sequence, we fit a Gaussian Process Regressor (GPR) to provide uncertainty measure over the predictions.

GPR assumes a kernel that determines the covariance over target functions and uses the observations to obtain a likelihood function. A new posterior distribution can be computed based on Bayes’ theorem. The choice of kernel characterized by a covariance function that measures the similarity between data points is an essential component in GPR. For the purpose of this study, this covariance function is chosen to be the popular Radial Basis Function (RBF). For more technical background on GPR, please refer to [165].

For each GPR corresponding to an edge, we use the observed samples $A_{\text{edge},t}$ to optimize the length-scale hyperparameter in the RBF kernel based on maximum-log-likelihood estimation. Using the fitting regressor function, a set of posterior samples up to the maximal forecast time horizon are predicted. Leveraging the probabilistic nature of GPR, we evaluate N samples drawn from the GPR at given inputs (in our case, a time step in the future). These N samples drawn from the Gaussian distribution at given t provide a range of values for the edge weight, and ultimately result in an uncertainty quantification of group membership (after aggregating edge forecasts into affinity matrix and application of the DS clustering N times).

Table 3.5: Performance of forecasting conversation groups at different future time steps for the Conflab dataset (cam 6). Note that the column $t = T$ indicates the detection results. Uncertainty quantified by standard deviation across all scenes are shown in parenthesis.

	$t = T$	$t = T+1$	$t = T+2$	$t = T+3$	$t = T+5$	$t = T+10$
F1 @ $Thr = 2/3$	0.90	0.88 (0.03)	0.86 (0.04)	0.84 (0.05)	0.80 (0.07)	0.76 (0.08)
F1 @ $Thr = 1$	0.76	0.73 (0.06)	0.69 (0.06)	0.69 (0.08)	0.66 (0.10)	0.62 (0.11)

From the validation sets of the data, we found that a sequence length of 10 was the optimal hyperparameter for affinity prediction and hence, we set $T = 10$ to acquire corresponding observed samples of affinity scores to fit the GPR for this forecasting task. The forecast horizon Z represents the time steps beyond T (measured in seconds for the Conflab dataset). Note that the fitted GPR could be sampled continuously; we chose a set of discrete time steps beyond T for the scope of this paper. Table 3.5 shows the forecasting results of predicting the conversation groups in Conflab (cam 6) using the aforementioned approach. We report the averaged F1 scores from evaluating the N affinity matrix instances

of each scene (i.e. aggregated from using the N samples of affinities drawn from each edge) for all scenes. The results show that there is a decreasing trend in the group detection performance as the forecast horizon extends, while the uncertainty in the group prediction in future scenes increases.

3.6 CONCLUSION AND FUTURE WORKS

In this work, we introduce and evaluate a deep learning joint-LSTM based neural network for pairwise affinity prediction, followed by DS clustering approach, for the task of conversation group detection in social settings such as cocktail parties and networking events. We motivate this LSTM-based approach to leverage the inherent temporal dynamics of human behaviors who could affect interactions and conversation groups. We showed that for the Conflab dataset (which has more temporal granularity compared to other existing datasets), our method shows improved performance in pairwise affinity predictions and therefore, leading to improved performance of conversation group detection. We further showed an analysis of the predicted affinity predictions and how they change overtime, which could be indicative of moments leading up to group formations and breaking. Lastly, we provide a forecasting framework based on our approach which predicts conversation groups at future time steps.

One of the limitations of this work include its performance in sparsely labelled data, such as for the Cocktail Party and SALSA dataset. The lower annotation frequency implies more varied conversation groups between time steps, and that the continuously changing group behavior in real life is not captured in the ground truth. Moreover, our use of ground truth features was partially motivated by what was provided in the existing datasets, and allows us to investigate the model performance without potential confounding sources of errors. However, this choice also does not shed light on the sensitivity of the performance with automatically acquired features, which would ultimately be more relevant in automatic systems (e.g., a social robot).

The model architecture could be further revised to take advantage of multimodal data at full sampling rates, for video, audio, and body movement motion. We note that the Conflab dataset contains manually annotated positions and orientations at 60Hz, as well as full 9 Degrees-of-Freedom IMU motion data captured at 56Hz sampling rate and speaking status annotations at 60Hz. While the trade-off among the difficulty of acquiring all of these data in an application setting, building and deploying a larger model, and the potential increase in performance should be considered, we believe that using more expressive modalities at finer temporal resolution, conversational group dynamics may be more thoroughly captured.

For further extension, the proposed forecasting framework presents an opportunity for researchers to detect individuals' intent to interact with others. More socially intelligent automated systems can be built if they are able to forecast affinities as a proxy for intention. Based on whether or not the predictions align with what actually occurs in the future, applications that are cognizant of what humans *plan* or *want* to do can be designed to enable better social interactions.

4

HEAD AND BODY ORIENTATION ESTIMATION WITH SPARSE WEAK LABELS IN FREE STANDING CONVERSATIONAL SETTINGS

4

This chapter is published as:

Stephanie Tan, David MJ Tax, and Hayley Hung. **Head and Body Orientation Estimation with Sparse Weak Labels in Free Standing Conversational Settings**. Understanding Social Behavior in Dyadic and Small Group Interactions. Proceedings of Machine Learning Research (PMLR). 2022.

4.1 INTRODUCTION

Studying social scenes that have free-standing conversation groups (FCGs) is of great interest. FCGs are a type of focused encounters that emerge in many social occasions, such as a cocktail party, a coffee break, a networking event, etc [6]. We find relevance in studying these social entities in order to study human interactions as part of the complex social dynamics. Prominent non-verbal cues that depict the social interplays are participant head and body orientations. With accurate estimations of head and body orientations, high-level social concepts such as conversation group formations and schisms can become more explainable.

Head and body orientations of participants are necessary prerequisites for many downstream tasks such as turn-taking patterns, conversation group memberships, estimation of social attention, etc [19]. Some tasks may only require either the head or body orientation. When identifying addresser/addressee or speaker/listener in conversations, head orientations are the primary cues [166]. When estimating group memberships, body orientations are the primary cues [23]. However, Langton et al. [16] have shown that head and body orientations are both important cues for estimating social attention. In social scenes such as Figure 4.1, eye gaze direction cannot be reliably observed; the attention target positions are not fixed throughout time, and the number of attention targets is not predefined. Under these adverse circumstances, attention direction is difficult to estimate. Hence, the importance of robust and accurate head and body orientations becomes more evident.

While there are many successes in human pose estimation and orientation estimation using deep learning frameworks [167–170], these methods only work well when human faces and body parts are easily discernible. Head and body orientation estimation remain challenging, especially for crowded scenes with relatively static subjects captured by videos from elevated side-views which result in low resolution, low light visibility, background clutter and occlusions (Figure 4.1 for example) [171]. In these settings, off-the-shelf deep learning methods are not effective [172] and retraining/finetuning them requires a considerable number of labelled samples. This motivates our proposed method under the transductive and few-labels setting which simultaneously estimates head and body orientations by leveraging wearable sensor data in addition to videos.

Recent advances have shown the efficacy in using a multi-view camera and multi-sensor scenario [28, 173]. The multi-view camera setting offers different viewpoints on people in



Figure 4.1: Examples of head and body orientation estimation challenges from the SALSA dataset [45] as highlighted in red: (a) low resolution, (b) low visibility, (c) background clutter, and (d) occlusion.

the scene for better acquisitions of head and body orientations. More interestingly, wearable sensors such as inertial measurement units (IMUs), microphones, infrared or Bluetooth proximity sensors, etc. have demonstrated an ability to recover subject orientations independently of the video modality [89, 174, 175]. In scenarios where video and microphone audio data are both recorded, a multimodal approach of head orientation estimation can be more accurate and robust than a unimodal one, as shown by Canton-Ferrer et al. [174]. Microphone data indicate who the speaker is at a given moment, and it is well known that the speaker tends to be the center of visual focus of a conversation group [176]. Considering these two aspects and given the ground positions of the interactants in free-standing scenarios, head orientations can be more reliably predicted in a complementary manner, especially when video data is partial or missing.

Despite the benefits that multimodal data from wearable sensors may offer, it is challenging to work with them. This is illustrated most evidently by the lack of in-the-wild datasets capturing natural interactions and emphasizing ecological validity in this domain, as it requires monumental effort to collect and annotate. Malfunctions of wearable sensors during data collection are more difficult to notice compared to those of video cameras. The types of different sensor noise are also hard to characterize. The resulting data could be of low quality, partial and/or missing due to periodic dropouts in sensor data streams, etc. [177, 178]. However, we argue that these difficulties are not reasons to deter from exploiting multimodal data from wearable sensors because the available data could still be of great value, as shown by literature [28, 173].

In this work, we highlight the possibility and advantage of working with a small number of human annotated orientation labels, along with sparse, noisy but automatically acquired labels from wearable sensors. As mentioned previously, wearable sensors are hard to work with. While it is possible to estimate labels from wearable sensors, the label quality varies depending on raw wearable sensor data quantity and quality. Hence we refer them as weak labels in this paper. Our results show that having information provided by other modalities like wearable sensors can indeed improve the performance of head and body orientation estimations in this free-standing conversation setting.

This study simultaneously addresses the following context where: 1) there is a relatively small number of head and body orientation samples ($\sim 10^2 - 10^3$) for each subject, 2) we jointly predict head and body orientation classification labels for unobserved samples only using a very small number ($\sim 5\%$) of sparsely distributed ground truth labels, 3) we take advantage of the temporal structure within the orientation label data and improve upon a previously suggested model based on Gaussian process regression (GPR) [88], and 4) most importantly, we fully exploit the utility of head and body orientation weak labels in addition to very few ground truths to improve performance.

4.2 RELATED WORK

4.2.1 HUMAN POSE ESTIMATION

Recent developments of deep learning methods [179–182] had greatly advanced 2D human pose estimation. Even though results are promising, addressing existing challenges such as low resolution and heavily occluded targets [172], and cluttered and crowded backgrounds, is an active research topic. Popular off-the-shelf pose estimation methods such as Openpose

[179] use a bottom-up approach to first detect body joints and later form associations to estimate a full skeleton model for each person in the frame. However, having only body part locations does not provide enough information to directly estimate the orientations of those body parts.

Using 3D pose estimation methods or converting 3D pose datasets [183] allows for extraction of orientations. Recent methods focusing on 3D pose estimations (full body, hand+body, etc.) [184–186] could be promising to directly infer orientations from 3D skeletal poses. However, orientation estimation could be decoupled and simplified from 3D pose estimation problem as there is evidence showing orientation estimations for objects can perform better when using 2D image features than 3D landmarks [187]. 3D poses may be difficult to infer due to occlusion or low resolution body parts, which are relevant scenarios in crowded social interactions in-the-wild. To address occlusion for 3D poses is an ongoing topic with [188] showing initial success on the MuPoTS dataset through localization and pose estimation with temporal smoothing.

4.2.2 HEAD AND BODY ORIENTATION ESTIMATION: RGB DATA

Previous works (e.g., [19, 189]) in head and body orientation estimation saw successes in using methods based on probabilistic frameworks (e.g. dynamic Bayesian networks, hidden Markov models, etc.). Taking advantage of the physical constraint of relative head and body pose and walking direction. Chen and Odobez [117] focus on the joint estimation of head and body orientation to achieve improved results. This body of work targets orientation estimations in a specific context by exploiting facial landmarks or motion priors; while this paper differentiates itself by focusing on the task in the surveillance setting with relatively static subjects. Without large movement towards one direction as a cue, orientation estimation becomes more difficult. Overall, there is more previous work on head orientation estimation compared to that of the body in the surveillance and crowded setting. In this particular context, human heads can be more easily seen and therefore head orientations are easier to predict. Human bodies can be occluded, making body orientations predictions more difficult. Lee et al. [190] proposed CRPNet that works well with low resolution images. However, their design goal favors speed over accuracy.

We acknowledge that there is a number of deep learning based methods [112, 114] for head and/or body orientation estimation problems. Most available methods are trained on datasets [191] that contain facial views. Applying Beyer et al.’s method [114] to SALSA is not straightforward because of the multi-camera setting and the extent of facial and body part occlusions. A body orientation estimation method proposed by Choi et al. [192] also faces similar challenges as head orientation estimation methods. Raza et al. [193] reported a joint head and body orientation estimation model using a hierarchical convolutional neural network. This pre-trained model trained with relatively small datasets (e.g., Human3.6M [183]) would most likely only be suitable for estimating orientations for pedestrians, and not for crowded and static social scenes like SALSA. Overall, the development of generalizable deep learning solutions for head and body orientation estimations are held back because of the lack of large scale datasets and the lack of environment/context variety in the training images. This constraint was only recently pointed out and addressed by the release of the COCO-MEOW dataset [194], which would enable future new data-intensive head and body orientation estimation methods.

Previous works [111, 114] showed that regression of head orientations can be achieved. Tasks such as predicting social attention [20] and personality traits [29] may benefit from more fine-grained orientation estimations. While regression is more descriptive, it is also challenging compared to orientation classification. As indicated by [173], the annotation noise for head orientation labels from video annotations is around 17° , which is more than the bin-width of class in our setting. Further experiments show that regression could be more advantageous, but the increase in performance on average is small compared to the variance. For the scope of this work, we reduce the orientation estimation problem to an 8-class classification problem (i.e., dividing 360° into eight sectors).

4.2.3 HEAD AND BODY ORIENTATION ESTIMATION: DEPTH AND WEARABLE SENSORS

Depth images can be used in estimated orientations. However, many works in this area (e.g., [195–197]) rely on the detection of the face and/or localization of facial and body landmarks. Works such as [198] combines RGB and depth data to estimate human body orientations. It is still challenging for subjects in crowded social scenes because of heavy occlusions with little motion cues.

In the sensor signal processing community, it is common practice to use wearable systems that house IMU sensors. To estimate orientations, IMU data serve as inputs to algorithms such as quaternion-based Extended Kalman Filtering and more recently reinforcement learning based methods [89, 175, 199, 200]. Ahmed and Tahir [201] showed that errors in estimating body part orientations while doing multi-axial actions such as waving and walking are generally as low as 2° . More recently, Webber and Rojas [202] showed the efficacy of using IMUs for human activity recognition without explicitly recovering the orientations (i.e., through gyroscopes). While IMU data could be valuable information for multimodal head and body orientation approaches, existing resources [28, 41] that focus on social scenes only contain accelerometer data, which is not enough for orientation recovery.

One approach to obtain estimations of head and body orientations is to use the proximity and audio information. Proximity sensors and microphones are already incorporated in the implementation of wearable badges that are common in the social signal processing and affective computing community (i.e., sociometric badges [61], OpenBadge [157], etc.). In turn, head and body orientations can be indirectly extracted. Previous work [45] used subject ground positions along with speaker/non-speaker correlations and proximity pings to estimate labels of head and body orientations, respectively. Compared to orientation labels from video or IMU, these estimated labels are less reliable since they are derived information from noisy sources. Nonetheless, they can still be explored and it is the focus of this paper.

4.3 OVERVIEW OF THE APPROACH

Our approach combines 4 kinds of inputs: 1) head and body visual features extracted from head and body image patches, 2) estimated head orientation labels from audio recordings, 3) estimated body orientation labels from infrared proximity sensors, and 4) manually annotated labels of some, but not all, frames. Note that the subject ground positions are

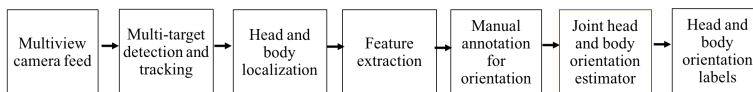


Figure 4.2: Overall work flow of automatic orientation estimation. The focus of this paper is outlined in red.

assumed to be given for acquiring inputs 2 and 3. The goal of this study is to jointly predict head and body orientations as an 8-class classification problem (dividing 360° into eight sectors) using matrix completion in a transductive learning setting. Matrix completion attempts to fill in missing entries in a matrix, which correspond to unobserved orientation labels. It is often solved by iterative optimization. Due to the sparsity and noise in the labels, the underlying challenge is to predict the head and body orientations which are temporally smooth. They also have to be consistent with the manual labels, weak labels (from wearable sensors), and the physical constraints that tend to couple the head and body behavior. For the purpose of this study, we consider multi-person tracking in videos, head and body detection, and appearance-based visual feature extraction as upstream tasks (Figure 4.2). The core of the proposed model (joint head and body orientation estimator in Figure 4.2) based on matrix completion is discussed in Section 4.4, followed by details on experimental conditions in Section 4.5.

4.4 PROPOSED MODEL

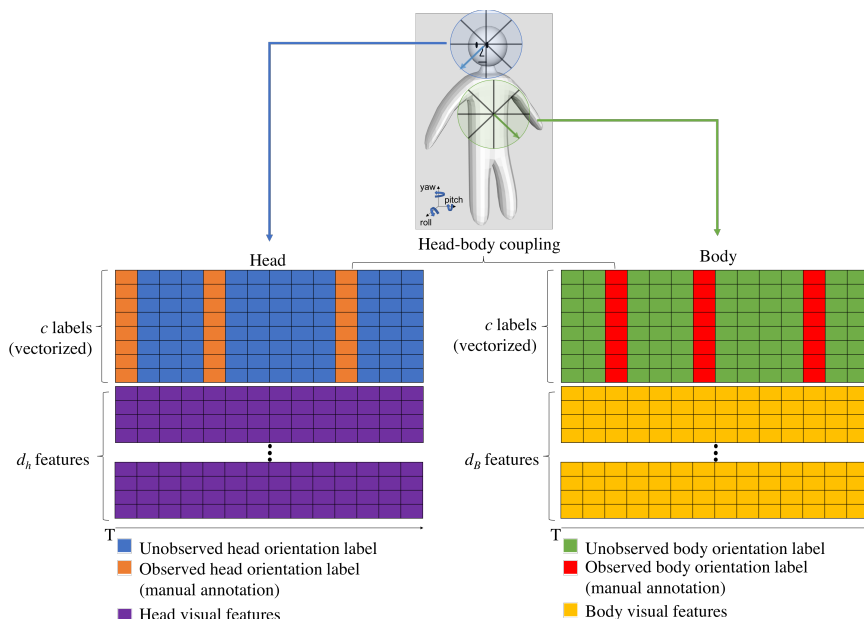


Figure 4.3: Graphical representation of the feature-label matrix. Head and body orientations are determined by 2D projections of yaw orientations.

In the supervised learning setting for a linear classifier, the objective is to learn the weight matrix $\mathbf{W} \in \mathbb{R}^{c \times (d+1)}$ by minimizing the loss on a training set $\mathcal{N}_{\text{train}}$ as

$$\arg \min_{\mathbf{W}} \sum_{i \in \mathcal{N}_{\text{train}}} \text{Loss} \left(\mathbf{Y}_i, \mathbf{W} \begin{bmatrix} \mathbf{X}_i \\ 1 \end{bmatrix} \right). \quad (4.1)$$

\mathbf{W} maps the d -dimensional features space $\mathbf{X} \in \mathbb{R}^{d \times T}$ to the c -dimensional (number of classes) output space $\mathbf{Y} \in \mathbb{R}^{c \times T}$ where T denotes the number of samples in time.

When dealing with noisy features and fuzzy labels, previous research [203–205] have empirically shown the practicality of casting a classification problem into a transductive learning setting such as matrix completion. For our specific task, borrowing from the linear classifier setting, a heterogeneous matrix is built by concatenating the orientation labels $\mathbf{Y} \in \mathbb{R}^{c \times T}$, visual features $\mathbf{X} \in \mathbb{R}^{d \times T}$, and a row of 1's (to model for bias) as

$$\mathbf{J} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \\ \mathbf{1} \end{bmatrix}, \quad (4.2)$$

where $\mathbf{J} \in \mathbb{R}^{(c+d+1) \times T}$.

Note that in (4.2), \mathbf{Y} is a vectorized one hot representation of orientation labels. Dividing 360° into eight sectors means that there are eight possible classes and each orientation belongs to one of the eight classes. For example, an angle θ that is $45^\circ \leq \theta < 90^\circ$ would be indicated by the vector $[0, 1, 0, 0, 0, 0, 0, 0]^\top \in \mathbb{R}^{c \times 1}$. Head and body label matrices are denoted by $\mathbf{Y}_h \in \mathbb{R}^{c \times T}$ and $\mathbf{Y}_b \in \mathbb{R}^{c \times T}$ respectively. The feature matrices $\mathbf{X}_h \in \mathbb{R}^{d_h \times T}$ and $\mathbf{X}_b \in \mathbb{R}^{d_b \times T}$ contain the visual features from head and body crops of each person, where d_h and d_b denote the respective feature dimensionality. Following the definition in (4.2), the visual features and corresponding labels are concatenated into two heterogeneous matrices $\mathbf{J}_h = [\mathbf{Y}_h^\top, \mathbf{X}_h^\top, \mathbf{1}^\top]^\top$ and $\mathbf{J}_b = [\mathbf{Y}_b^\top, \mathbf{X}_b^\top, \mathbf{1}^\top]^\top$ for head and body orientation estimation respectively (Figure 4.3). In addition, a projection matrix $\mathbf{P}_h = [\mathbf{I}^{c \times c \times T}, \mathbf{0}^{c \times T \times (d_h+1) \times T}]$ is introduced to extract only the head orientation labels from the heterogeneous matrix \mathbf{J}_h . In a similar manner, a projection matrix $\mathbf{P}_b = [\mathbf{I}^{c \times c \times T}, \mathbf{0}^{c \times T \times (d_b+1) \times T}]$ is defined to extract body orientation labels.

The unobserved orientation labels can either be initialized by information provided by external sources or simply set to zero. In this study, we take the first option. The initial matrices for head and body orientations are denoted by $\mathbf{J}_{0,h}$ and $\mathbf{J}_{0,b}$ respectively. The label matrix in $\mathbf{J}_{0,h}$, denoted by \mathbf{Y}_h , is further divided into a training set $\mathbf{Y}_{\text{train},h}$ and a test set $\mathbf{Y}_{\text{test},h}$. Similarly, the label matrix in $\mathbf{J}_{0,b}$, denoted by \mathbf{Y}_b , is divided into $\mathbf{Y}_{\text{train},b}$ and $\mathbf{Y}_{\text{test},b}$. Each training set consists of observed labels, while the test set consists of labels to be predicted. We assume that the training and test set samples are interleaved, as shown in Figure 4.3. We chose this assumption because this could be reflective of real-life scenarios of having observed and unobserved samples intermittently. For the sake of brevity, the subsequent discussion focuses on the head orientation matrix. The body orientation matrix and its corresponding optimization formulation are analogous.

The following discussion outlines the proposed matrix completion method based on the aforementioned setting. We formulate it as an optimization problem, consisting of four components: 1) enforcement of feature-label linear dependency, 2) temporal smoothing, 3) regularization by weak labels, and 4) head-body coupling. Each component applies

to completing matrices for estimating head and body orientation respectively. The joint completion of the head and body matrices are further explained in Section 4.4.5.

4.4.1 RANK MINIMIZATION

Following the linear classifier assumption from (4.2), previous work [205] has shown that the matrix \mathbf{J}_h should be low rank. The linear classifier in (4.1) requires that there is row dependency in (4.2), hence low rank. The objective is to recover the missing orientation labels such that the rank of the heterogeneous matrix \mathbf{J}_h is minimized. Rank minimization is a non-convex problem [205]. However, Candes and Tao [206] showed that $\text{rank}(\mathbf{J}_h)$ can be relaxed to its tightest convex envelope which is the nuclear norm, $\|\mathbf{J}_h\|_*$, i.e.

$$\text{rank}(\mathbf{J}_h) \approx \|\mathbf{J}_h\|_*. \quad (4.3)$$

In practice, the optimization problem then becomes a minimization of the nuclear norm of \mathbf{J}_h .

4.4.2 TEMPORAL SMOOTHING

If samples in the heterogeneous matrix are temporally sorted, one can take advantage of the temporal structure between the columns. Orientation labels are, to an extent, temporally smooth, as head and body poses are not expected to change drastically within a short time period. This can be seen as a column-wise regularization. An interpolated time series of orientation labels $\tilde{\mathbf{Y}}_h$ can be generated using an appropriate interpolation scheme to estimate the unobserved orientation labels. In the proposed method, Gaussian process regression (GPR) is chosen as the interpolation scheme. Also known as Kriging, GPR has the same objective as other regression methods, which is to predict the value of a function at some point using a combination of observed values at other points. Rather than curve fitting using a polynomial function for instance, GPR assumes an underlying random process, more specifically a Gaussian process [207], from which the observed values are sampled. A new posterior distribution is computed based on the assumed (Gaussian process) prior and Gaussian likelihood functions [208]. The Gaussian process prior is characterized by a covariance function which measures the similarity between data points; and thus the choice of a suitable covariance function is an essential component in GPR. More details of Gaussian processes and Kriging can be found in [209].

Following this procedure, we denote $\mathbf{Y}_{\text{GP},h} \in \mathbb{R}^{c \times T}$ as the label matrix where the missing values are imputed by the prediction of GPR. After acquiring the interpolated labels, a new matrix $\mathbf{J}_{\text{GP},h}$ is defined as

$$\mathbf{J}_{\text{GP},h} = \begin{bmatrix} \mathbf{Y}_{\text{GP},h} \\ \mathbf{X}_h \\ \mathbf{1} \end{bmatrix}. \quad (4.4)$$

We introduce an additional squared loss term $\|\mathbf{P}_h(\mathbf{J}_h - \mathbf{J}_{\text{GP},h})\|_F^2$ to the optimization problem, where $\|\cdot\|_F^2$ is the Frobenius norm. It is a regularization to ensure that the predicted labels do not deviate drastically from those obtained using temporal interpolation. The projection matrix \mathbf{P}_h ensures that the loss is only considered over the orientation labels.

Note that GPR is an example of a regression method that works well in this setting. Alternative regression methods such as Laplacian smoothing [45], piece-wise linear inter-

polation and polynomial regression can also be applied. Our justification for this choice is presented in Section 4.6.

GAUSSIAN PROCESS REGRESSION KERNELS

The basis of GPR is Gaussian Process (GP). A GP is defined to be a random process $f(t)$ for $t \in T$, such that for every finite subset of selected time steps $\{t_1, t_2, \dots, t_N\}$, $\{f(t_i); i = 1, 2, \dots, N\}$ is jointly normally distributed. A GP is necessarily defined by its mean function $m(t) = \mathbb{E}[f(t)]$ and its covariance function, also called kernels, $k[t_i, t_j] = \mathbb{E}[(f(t_i) - m(t_i))(f(t_j) - m(t_j))]$. While the mean function is often chosen to be zero, the choice of kernels in GP Regression is critical, and is known to affect performance to a great extent. It controls the degree to which data are smoothed when estimating the unknown function [210]. In GP, the kernel represents distance or similarity between two latent variables $f(t_i)$ and $f(t_j)$ given inputs t_i and t_j , $i \neq j$. Intuitively, It describes how output $f(t_j)$ can be affected by output $f(t_i)$. There are many options for these kernel functions. The radial basis function (RBF) kernel is most commonly used and is represented as follows

$$k(t_i, t_j) = \sigma_f^2 e^{-\frac{1}{2} \frac{(t_i - t_j)^2}{\sigma_l^2}}, \quad (4.5)$$

where σ_l denotes the characteristic length scale that controls the smoothness of the function and σ_f determines the vertical variation.

Matérn kernels are a class of kernels that provide extra flexibility compared to the RBF kernels in controlling the differentiability of the sample functions drawn from the GP distribution. Matérn kernels are of the form

$$k(t_i, t_j) = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} |t_i - t_j|}{\sigma_l} \right)^\nu K_\nu \left(\frac{\sqrt{2\nu} |t_i - t_j|}{\sigma_l} \right) \quad (4.6)$$

where ν is the differentiability parameter and K_ν is the modified Bessel function of the second kind. Sample functions drawn from GP with Matérn kernels are $(\nu - 1)$ times differentiable, whereas a GP with RBF kernels lead to sample functions that are infinitely differentiable. The parameter ν is usually chosen to be $\frac{3}{2}$ or $\frac{5}{2}$, and (4.6) can be simplified, respectively, as

$$K_{\frac{3}{2}}(t_i, t_j) = \sigma_f^2 \left(1 + \frac{\sqrt{3}(t_i - t_j)}{\sigma_l} \right) e^{-\frac{\sqrt{3}(t_i - t_j)}{\sigma_l}}, \quad (4.7)$$

and

$$K_{\frac{5}{2}}(t_i, t_j) = \sigma_f^2 \left(1 + \frac{\sqrt{5}(t_i - t_j)}{\sigma_l} + \frac{5(t_i - t_j)^2}{3\sigma_l^2} \right) e^{-\frac{\sqrt{5}(t_i - t_j)}{\sigma_l}}. \quad (4.8)$$

The kernels in (4.7) and (4.8) lead to once and twice differentiable sample functions in GP. In the subsequent discussion, we refer to (4.7) and (4.8) as Matérn 3/2 kernel and Matérn 5/2 kernel, respectively.

The GP kernel function is often chosen based on a qualitative understanding of the underlying data [211]. Though RBF kernels are most commonly used, it has been shown that Matérn kernels are more suitable to model physical processes [212, 213]. Sample functions tend to be less smooth when using Matérn kernels due to finite differentiability, allowing for more realistic capturing of the process. In the context of head and body orientations, it is unlikely that the unknown function would be very highly differentiable.

4.4.3 REGULARIZATION BY WEAK LABELS

Weak labels estimated from sources such as wearable sensors could be informative though they might be less precise than ground truth (GT) labels. They could still provide additional information that assists in the classification task. We propose a regularization term that incorporates weak labels of head and body orientation. The regularization term can be written as

$$\|\mathbf{P}_{w,h}(\mathbf{J}_h - \mathbf{J}_{w,h})\|_F^2, \quad (4.9)$$

where $\mathbf{P}_{w,h}$ is a projection map that extracts the portions where weak label readings are available. The formulation of $\mathbf{J}_{w,h}$ is analogous to (4.4), where weak labels are treated as approximations of the actual labels. Note that multiple regularization terms of the same form as (4.9) can be added to the formulation depending on the number of weak labels sources. This highlights the flexibility and modularity of the proposed model in the context of multimodal head and body orientation estimation.

4.4.4 HEAD AND BODY COUPLING

Previous research [45, 117, 214] has shown that coupling head and body orientation estimation is advantageous for improving accuracy. The proposed formulation also captures the physical constraints between head and body orientations. Since head and body orientations are jointly estimated, this relation fits in nicely as an additional regularization to the optimization problem. It is reasonable to model that head and body orientations cannot be too different at any given time step. Though hinge loss would probably be more appropriate, the relation can also be captured by squared loss, for the ease of analytical derivation and numerical optimization. The regularization term can therefore be written as $\|\mathbf{P}_h \mathbf{J}_h - \mathbf{P}_b \mathbf{J}_b\|_F^2$.

4.4.5 OPTIMIZATION PROBLEM

To summarize, the entire optimization problem, considering all the regularizations and indicating terms associated with both head and body (described in Sections 4.1-4.3), is given by

$$\begin{aligned} & \mathbf{J}_h^*, \mathbf{J}_b^* \\ &= \arg \min_{\mathbf{J}_h, \mathbf{J}_b} \underbrace{\nu_h \|\mathbf{J}_h\|_* + \nu_b \|\mathbf{J}_b\|_*}_{\text{matrix low-rankedness}} \\ & \quad + \underbrace{\frac{\lambda_h}{2} \|\mathbf{P}_h(\mathbf{J}_h - \mathbf{J}_{GP,h})\|_F^2 + \frac{\lambda_b}{2} \|\mathbf{P}_b(\mathbf{J}_b - \mathbf{J}_{GP,b})\|_F^2}_{\text{temporal smoothing}} \\ & \quad + \underbrace{\frac{\gamma_h}{2} \|\mathbf{P}_{w,h}(\mathbf{J}_h - \mathbf{J}_{w,h})\|_F^2 + \frac{\gamma_b}{2} \|\mathbf{P}_{w,b}(\mathbf{J}_b - \mathbf{J}_{w,b})\|_F^2}_{\text{weak label regularization}} \\ & \quad + \underbrace{\frac{\mu}{2} \|\mathbf{P}_h \mathbf{J}_h - \mathbf{P}_b \mathbf{J}_b\|_F^2}_{\text{head-body coupling}}, \end{aligned} \quad (4.10)$$

where v_h , v_b , λ_h , λ_b , γ_h , γ_b and μ are weights that control the trade-off between the different terms. The equation in (4.10) can be solved iteratively by an adapted Alternating Direction Method of Multipliers (ADMM) [45, 215] to jointly solve the minimization problem for the head and body orientation matrices.

Derivation and implementation details are included in Appendix 7.5. Note that an advantage of the weak labels regularization is that we don't need to study in great detail the quality of the weak labels beforehand. Hyper-parameter optimization will determine the coefficients such that high quality weak labels boost the performance and low quality weak labels get disregarded automatically in squared loss term in (4.9).

4.5 EXPERIMENTS

4

This section provides a brief introduction of the SALSA dataset [45] that was used to obtain the experimental results, and an overview of the experimental protocol. Note that since the premise of our learning problem is transductive and we target a setting with very small number of training samples and labels as well as using multimodal data, we do not compare our method to existing deep learning methods (for head and body orientation estimation) which rely on (re)training on much larger number of labeled data that contain images only and are not multimodal.

4.5.1 SALSA DATASET ANALYSIS

SUMMARY

The SALSA dataset is a multimodal dataset that was captured at a social event that consists of a poster presentation session and a mingling event afterwards, involving 18 participants. For this study, we focus on the video recordings, proximity sensor pings, and audio data of the poster presentation session (~17 minutes). Ground truth labels of head and body orientation of each participant were manually annotated every 3 seconds. Additional details on annotations can be found in [45]. Head and body orientations were extracted from audio and proximity data respectively, independent of the video [45]. These are treated as weak labels in our context.

The SALSA dataset is a challenging dataset for head and body orientation estimation due to the low resolution of targets, cluttered background, and occlusions. The class distribution of the GT labels is shown in Figure 4.4. We discretized the GT labels, which are labeled with respect to the ground plane, into 8 angular bins $[0,45)$, $[45,90)$, $[90,135)$, $[135,180)$, $[180,225)$, $[225,270)$, $[270,315)$, and $[315,360)$ in degrees in the room coordinate system; and labeled serially from class 1 to class 8. The majority of GT labels correspond to non-frontal views of the subjects, hence making it difficult to estimate head and body orientations [86]. Overall, the dataset is relatively balanced except for class 5 and 6. However, person-wise data among the 18 subjects could be heavily imbalanced.

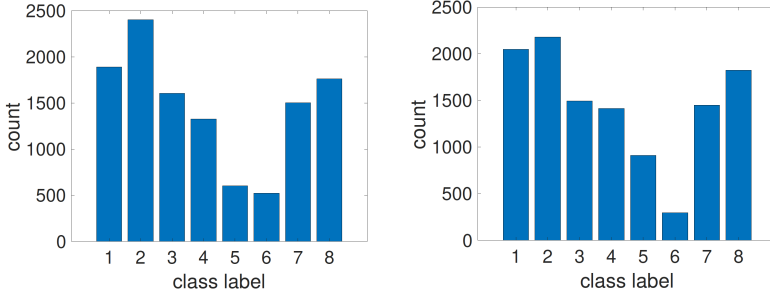


Figure 4.4: Overall class distribution of head (left) and body (right) GT labels in the SALSA dataset.

4

SALSA WEAK LABELS ANALYSIS

The weak labels estimated for each subject during the poster session of the SALSA dataset are sparse and/or noisy. Head orientation weak labels are extracted by correlating the speaking status between subjects. Body orientation weak labels are extracted based on proximity pings. Both procedures rely on the ground position and relative proximity of the subjects. Weak labels are 28% and 87% sparse for head and body, respectively. Hence, body orientation weak labels are unavailable for most of the poster session. The reason for weak label absence is unclear.

To quantify the quality of the available weak labels, we calculate the difference between the weak labels and GT labels. Since angles are periodic (i.e. repeat every 360°), we take the circular difference δ between the two discretized sets of labels

$$\delta = \min(|G_i - W_i|, N - |G_i - W_i|), \quad (4.11)$$

where G_i denotes the i^{th} GT label, W_i the i^{th} weak label, and N the total number of possible classes, which is 8 in the context of this paper. Therefore, the maximal difference does not exceed 4. If the difference is 0, then weak labels match with the GT labels. The distribution of differences in orientation labels are shown via histograms in Figure 4.5. As illustrated in the class difference distribution plots, there is generally a considerable discrepancy of class difference of 2 or 3 classes between weak head labels and ground truth. This is expected because microphone data are generally noisy, which can cause errors in estimating speaker and listener status. On the other hand, the class difference in body labels concentrated at 0 is obtained after adding 180° to all weak body labels. This is an artifact that has not been explicitly stated in the original SALSA dataset paper [45].

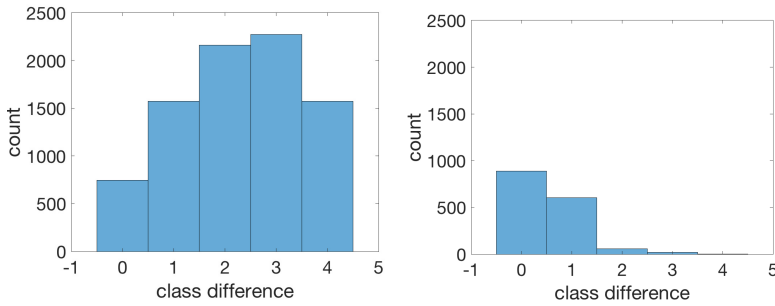


Figure 4.5: Distribution of class difference between ground truth and weak labels for head (left) and body (right) orientations.

Poster sessions include moments of high crowd density which compromises the quality of these weak labels, as auditory signals are cross-contaminated and infrared sensors may pick up pings from multiple directions in the vicinity. Previous work [45] considered weak labels to be the same quality as GT labels whenever they are available. Also another previous work [88] considered head and body orientation estimation as an isolated problem based on only video data. Unlike the aforementioned previous works, this paper exploits the potentially useful information provided by available weak labels. The regularization term in the formulation allows us to circumvent the associated intrinsic noisiness and sparsity (Section 4.4.3). We also report some investigatory results by simulating labels of different qualities and show how incorporating them via regularization can enhance the model performance. The purpose of this exercise is to provide further insight into future multimodal orientation estimation approaches.

4

4.5.2 EXPERIMENTAL SETUP

We used the Histogram of Gradients (HOG) visual features for head and body crops of each participant from the SALSA dataset poster session, which aligns with the choice in [45]. Similar to the approach proposed by Alameda-Pineda et al. [45], visual features from the four cameras are concatenated and Principle Component Analysis (PCA) was performed to keep 90% of the variance as dimensionality reduction preprocessing. This results in a 100-dimensional feature vector. Training data are the observed labels and test data are the unobserved labels to be predicted. In a transductive learning setting, since the objective is to predict labels for the unobserved entries only and not generalize to further unseen data, weights are not explicitly learned. Training data and test data partitions are determined by random sampling of columns (over time). Because of this randomness, training and test data are interleaved and we take advantage of this inherent structure in our formulation.

Previously, a person specific training and test scheme, in which a model is trained for every subject, was presented in [88]. A caveat of performance from this type of scheme is that there can be large inter-subject variation. The model trained on one subject may not generalize to other subjects. To investigate the generalizability of the proposed model in this paper, we introduce a person independent training/test protocol. Due to the small subject-wise sample size (18 subjects) of the SALSA dataset, we use a nested leave-one-person-out cross validation (LOPOCV) protocol to conduct the experiments. One subject is left out for each test fold, resulting in 18 folds overall. Within each training fold of

17 subjects, we use a 3-fold cross validation to select the hyperparameters (via Bayesian optimization) in the optimization problem (4.10). For each subject, the head and body orientation samples are arranged temporally and a random fraction of them are chosen to be training samples. Due to the randomness in this step, we repeat the process of randomly selecting the training samples five times within each of the three folds. We use Bayesian optimization to identify the hyperparameters with the negative of the sum of body and head orientation estimation classification accuracy averaged across the 17 subjects as the objective function.

The model performance on the test subject from each LOPOCV fold is evaluated using the best set of hyperparameters and averaged results from 18 folds are reported. For experimental conditions investigating the influence of the model parameters (Section 4.6), the model is retrained using the same protocol.

4

4.6 MODEL ANALYSIS

A comprehensive model analysis is conducted considering various possibilities in training schemes, kernel options, and a combination of regularization terms.

4.6.1 RESULTS

Table 4.1 reports two sets of baseline results along with results obtained from the proposed model trained using LOPOCV. To obtain the first naive baseline, we simply set the unobserved samples to the value of the mode of the selected samples. The second baseline is the set of person specific results which is reported in [88]. Table 4.1 shows the averaged-across-subject head and body orientation estimation results for different fractions of manual annotations. There is a notable increase in performance for the proposed model with respect to the two baselines. We also report performance of the proposed model without using the regularization based on the weak labels and observe that including the weak labels has a positive contribution to the performance.

The hyperparameters in the proposed model (4.10) are $\{v_h, v_b, \lambda_h, \lambda_b, \gamma_h, \gamma_b, \mu\}$. We arbitrarily set $v_b = 1$ as the contribution of the other terms can be considered relative to v_b . At 5% manual labels, hyperparameter optimization yields $v_h = 7.4$, $\lambda_h = 6.4$, $\lambda_b = 5.6$, $\gamma_h = 1.7$, $\gamma_b = 1.3$, and $\mu = 5.2$ averaged across 18 folds of LOPOCV. Comparing v_h and v_b , the low rankness of J_h carries more weight than that of J_b in (4.10). This corroborates the intuition that there is considerable occlusion of subjects' body and less occlusion of subjects' head. We also note that temporal smoothing in both head and body orientations (λ_h and λ_b), and head-body coupling (μ) are important to model performance.

Figure 4.6 shows a detailed subject-wise comparison at 5% manual labels (i.e., observed samples) fraction. For the majority of the subjects, we notice a consistent improvement with respect to the two baselines. Improvement with respect to the results from [88] is attributed to the optimization of the GP kernel and weak label regularization which were not considered previously. For some subjects such as subject 2 and 8, the mode baseline already performs well, especially for body orientation estimation. This is because orientation variation and diversity are relatively low for these subjects. Larger orientation diversity can lead to lower performance and higher variation across subjects [88]. On a higher level, this can be related to the personality and role functions of subjects, the

Table 4.1: Averaged classification accuracy (%) for different fractions (%) of manual annotations. Standard deviation (%) in accuracy performance across all people (in the LOPOCV framework) is shown in the parenthesis. State-of-the-art performance [45] at 5% manual annotation is 56.7% and 59.7% for head and body, respectively.

	Fraction	Mode Baseline	Tan et al. [88]	Ours	
				no weak labels	weak labels
Head	5	40 (13)	63 (13)	64 (13)	65 (13)
	30	41 (13)	68 (13)	72 (13)	74 (13)
	50	41 (13)	70 (13)	77 (9)	76 (12)
	70	41 (13)	71 (13)	77 (11)	77 (12)
Body	5	45 (18)	70 (13)	72 (13)	76 (12)
	30	47 (17)	79 (11)	81 (11)	83 (9)
	50	47 (17)	81 (10)	85 (11)	86 (9)
	70	47 (17)	83 (10)	86 (9)	86 (9)

dynamics between subjects, and the context of the social scene. For the other manual label fractions, the observations are similar.

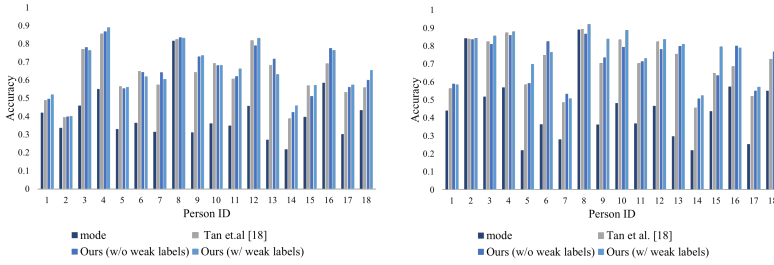


Figure 4.6: Comparisons of head (left) and body (right) orientation estimation at 5% manual annotation across four setups: mode baseline, Tan et al. [88], and our formulation without and with regularization by weak labels. The plots are best viewed in color.

4.6.2 KERNEL CHOICE

The choice of the kernel is a critical decision during the modeling process of GPR. Kernel functions encode the underlying behavior of the data such as its periodicity and smoothness. Since we are working with head and body orientation angles, the important feature to take into account is the smoothness. Even though head and body turns could be seen as smooth in general, we hope to capture sudden head and body turns which are more interesting for social scene analysis.

We focus on choosing among the RBF, Matérn 3/2 and Matérn 5/2 kernels. During the hyperparameter optimization, the Matérn 3/2 kernel was found to be the optimal option for all the different fractions of manual annotations listed in Table 4.1. It further supports with the assumption that head and body orientations are only mildly smooth over time. The RBF kernel assumes that the learned smoothing function is infinitely differentiable which doesn't appear to be as fitting in this particular modeling process. Similarly, the Matérn 5/2 kernel is twice differentiable while the Matérn 3/2 kernel is once differentiable. Further optimization of kernel parameters pertaining to the Matérn 3/2 kernel option was

also performed. Signal variance σ_f is a scaling factor that describes the variation of the regressed values to their mean. Characteristic length scale σ_l describes the smoothness of the function. The averaged hyperparameters σ_f and σ_l are 4.6 and 45 respectively.

4.6.3 REGULARIZATION BY WEAK LABELS

In this section, we discuss model performance with two different kinds of weak label inputs for the regularization term in (4.9). These inputs are used to populate the label portion of $\mathbf{J}_{w,h}$ and $\mathbf{J}_{w,b}$. First, we use the weak labels provided in the SALSA dataset. Despite the issues with the quality of weak labels as explained in Section 4.5.1, we include the results for instructive purposes. If a weak label is not available at a given timestep, we use the nearest available weak label in time.

4

The second kind of weak label inputs is artificially generated. We want to investigate how the performance changes with the quality of weak labels. To simulate a set of noisy weak labels, we generate a set of artificial labels by perturbing the GT labels. In practice, we add Gaussian noise with zero mean and standard deviation equal to 15, 30, 60, 90 and 120 degrees. This set of artificial weak labels acts in place of the actual weak labels from SALSA.

In Figure 4.7, we report the results obtained with these two types of weak labels. Artificial weak labels have been created with Gaussian noise of standard deviation equal to 30 degrees. The baseline model represents the case when no weak labels are included. We observe that using true weak labels decreases the performance compared to the baseline. This is expected given the poor quality of the actual weak labels. However, with artificial weak labels, there is a notable increase compared to the baseline. This shows that weak labels of decent quality can be exploited, especially when the manual annotation fraction is low. With an increasing number of observed samples, the number of unobserved samples becomes fewer, reducing the dependence on weak labels. As a result, the value of using weak labels diminishes with an increasing number of observed samples. But as we are especially interested in the regime of few observed samples, we highlight the fact that weak labels can indeed boost model performance.

Figure 4.8 shows the improvement in performance due to noisy weak labels with respect to the baseline model. We set 5% of the data as manual annotations or observed samples. When weak labels become increasingly noisy, the model performance falls below that of the baseline. This demonstrates that weak labels need to be of a reasonable quality to contribute positively to performance and justifies the poor performance when the true SALSA weak labels are included. Furthermore, the improvements in body orientation estimations are more consistent compared to those of the head. Hence, we emphasize that head orientation estimation is a more difficult task, possibly because head orientations vary more than body orientations over short time scales. A different approach such as classification with finer granularities could be promising for better head orientation estimations.

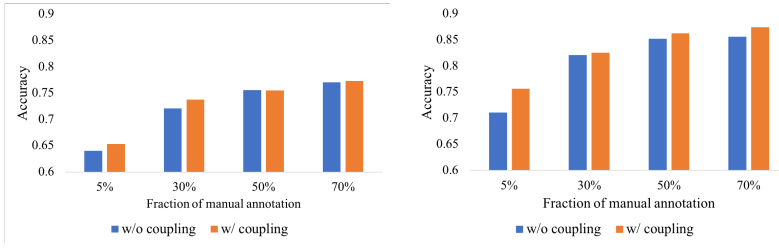


Figure 4.9: Performance comparison for head (left) and body (right) orientation estimation without and with head-body coupling regularization.

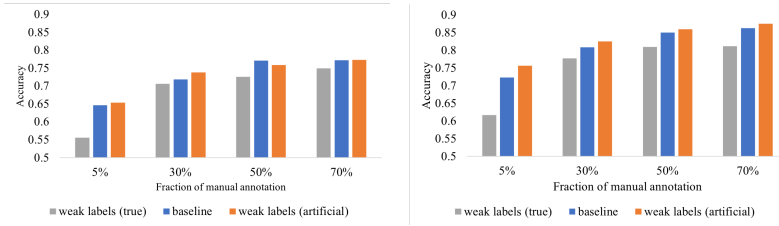


Figure 4.7: Performance comparison for head (left) and body (right) orientation estimation without (baseline) and with weak label regularizations. Artificial weak labels have been created using Gaussian noise of standard deviation equal to 30 degrees.

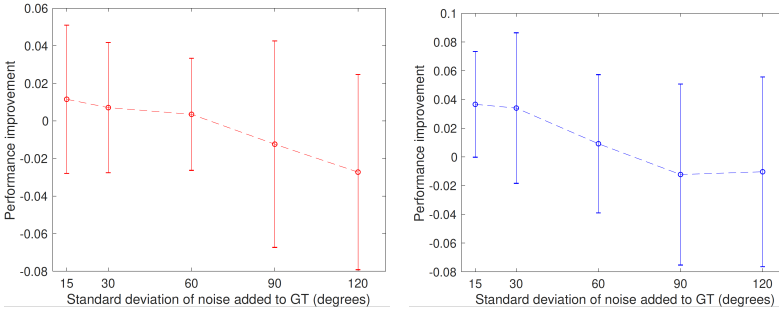


Figure 4.8: Improvement in performance of head (left) and body (right) orientation estimation for different magnitudes of noise in artificially generated weak labels. The improvement is reported with respect to the baseline (i.e., no weak labels) in the 5% observed samples setting. The error bars indicate subject wise standard deviation in improvement.

4.6.4 CONTRIBUTION OF HEAD-BODY COUPLING

To study the contribution from head-body coupling regularization term, we remove this from the best model (i.e. with artificial weak labels) and compare the performance difference. Figure 4.9 shows the extent to which the performance decreases without head-body coupling, which is more prominent when the manual annotation fraction is low. Similar to the observation made for the weak label regularization, when there is more observed samples, GP smoothing becomes advantageous and dominant, making the head-body coupling term less important. However, when the observed sample size is small, the head-body

coupling contributes positively to the performance. In particular, the effect is prominent in body orientation estimation where an increase of 4.7% in accuracy is obtained when 5% of the data is manually labeled.

4.7 DISCUSSION AND CONCLUSION

In this paper, we present a model that utilizes few labeled samples to classify unlabeled samples for head and body orientation estimation in a transductive setting using matrix completion. The formulation of the model combines rank minimization of the joint feature-label matrix, temporal smoothing over labels (based on GPR), weak labels regularization that takes advantage of weak labels from wearable sensors, and head-body coupling to ensure physical restraints of head and body orientation estimates. Since we are especially interested in investigating multimodal orientation estimation, we primarily test our method on the challenging SALSA dataset. SALSA is the largest annotated dataset that contains multiple overlapping video recordings and wearable sensor readings along with ground positions, and head and body orientations of each subject. In Section 4.5.1, we describe some issues and challenges with working with weak labels acquired from wearable sensors. We do not compare to existing deep learning methods for head and/or body orientation estimation (e.g., [112, 114, 192, 193]) because of the fundamental difference in learning setting and the lack of multimodal comparisons. Future extension of studies based on deep learning approaches could be developed to accommodate multimodal data for this task, upon further ablation studies to verify the efficiency of wearable sensing data.

Notable conclusions from our experimental results are – (i) the person independent model achieved by the proposed formulation outperforms the person specific model, which shows promising generalization ability; (ii) a more suitable kernel for GPR when modeling head and body orientation series is the Matérn 3/2 kernel, as opposed to the more popular RBF kernel; (iii) weak labels of low quality may impair performance but in the case where better quality weak labels are used, model performance is boosted; and (iv) head and body coupling indeed improves head and body orientation. The increase in performance due to (iii) and (iv) is especially notable in the few manual annotations or observed samples regime.

There are some limitations to this model. The performance would depend on the spacing (availability) of observed samples in order for temporal smoothing to be effective. The method does not apply to independent and isolated unseen samples. It would not perform well if the period of interest is far away in time compared to the observed samples. On the other hand, this provides initial guidelines on selecting which samples to annotate if there are financial constraints. Performance would also depend on the methods applied to the sensor signals as acquiring head and body orientation estimates from wearable sensors is challenging in itself.

Future work entails addressing the aforementioned limitations. On the other hand, given the flexibility of the model, possible topics to explore include but are not limited to matrix completion with missing features, feature representation across different modalities, and joint head and body matrix completion of several subjects, given prior information such as group membership assignments. In the case of a large number of unlabeled samples in a dataset, results from the proposed model would give competitive rough estimates of the actual labels as a data augmentation technique. This is a viable option if obtaining

manual labels becomes expensive or impossible. Acquiring results from the model is relatively computationally inexpensive, and we can use them as a springboard for deep neural networks or other models that require a larger number of labeled samples to achieve better head and body orientation estimations.

5

A MODULAR APPROACH FOR SYNCHRONIZED WIRELESS MULTIMODAL MULTISENSOR DATA ACQUISITION IN HIGHLY DYNAMIC SOCIAL SETTINGS

5

This chapter is published as:

Chirag Raman*, Stephanie Tan*, and Hayley Hung. **A modular approach for synchronized wireless multimodal multisensor data acquisition in highly dynamic social settings**. Proceedings of the 28th ACM International Conference on Multimedia. 2020.

*: equal contribution



Figure 5.1: A typical in-the-wild social interaction setting; adapted from the MatchNMingle Dataset [68]

5.1 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 [216, 217], or synchrony between participants [218]. 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 sometimes also 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 [66, 219, 220]. Concurrently, the ubiquitous computing community were developing approaches using wearable sensors that allowed for more pervasive sensing of social behaviors [221–223] 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 [45, 68]. Figure 5.1 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 [223].

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 Sec. 5.3.3 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

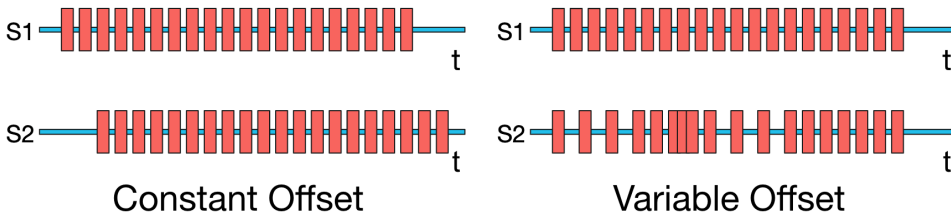


Figure 5.2: Basic types of desynchronization

is often timestamped in the frame-based SMPTE timecode format such as linear time code (LTC)- HH:MM:SS:FF [224]. 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 [225], 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) [226] 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 5.2 we review data recording or post-processing techniques used in other human behavior research and discuss the trade-offs involved. In Section 5.3 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 5.4 before discussing cost versus latency considerations in Section 5.5. Finally, we summarize our findings in Section 5.6.

5.2 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 5.2 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.

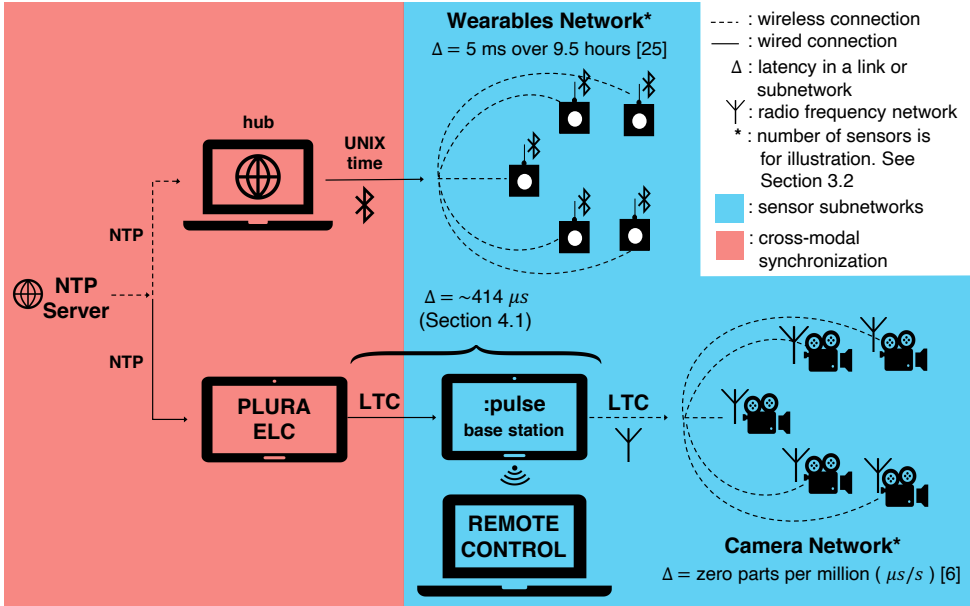


Figure 5.3: Overview of our proposed architecture. The reference time signal originates from the chosen NTP server and propagates to the subnetworks of wearable sensors and cameras.

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 [45]. 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 [227]. 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 [228]. In contrast, we propose a modular approach that synchronizes the devices at data acquisition, requiring minimal—if any—post processing for synchronization.

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 [24, 81, 229, 230]. 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 5.3.3.

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 [231] to analyze interest and affiliation in conference attendees [232]. 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 [233]. 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, [222] 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 [230]. 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 [219, 220]. 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 [234], although this only works for the video modality.

Lichtenauer et al. [66] 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 [66]. This approach was used in the recording of the MAHNOB-HCI datasets [235]. 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 [236], or in post-processing by performing an optimization over or manually annotated markers in a subset of frames [237]. 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 or 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 [68], 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 [62]. 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.

5.3 OUR APPROACH

5

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 5.3. 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 5.3.1. We describe a particular real-world instantiation of our system in Section 5.3.2, 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 5.3.3, which frames our subsequent experimental design.

5.3.1 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 [238] 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*;

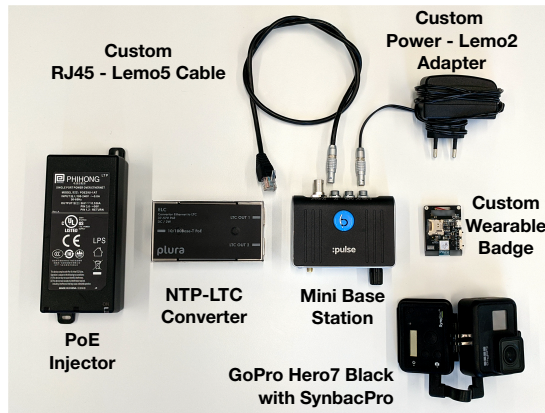


Figure 5.4: Core components in our setup depicted only with custom cables; the connectors are aligned with the corresponding sockets.

5

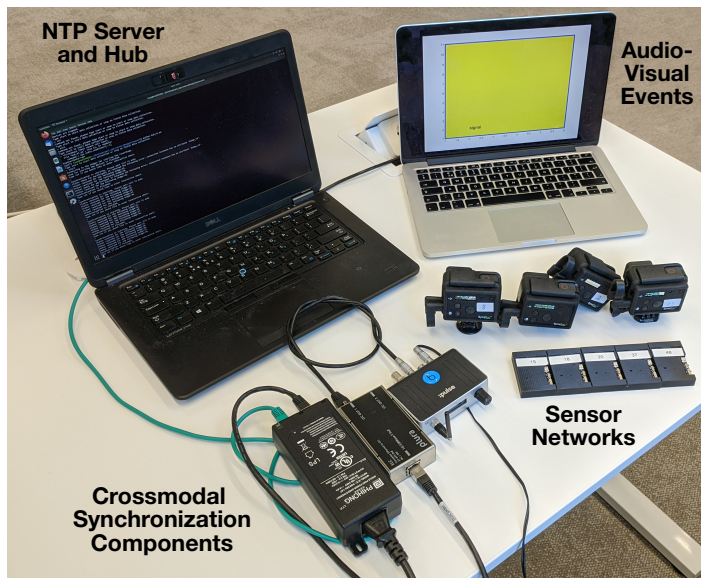


Figure 5.5: Full working setup of our data acquisition system, here shown with four cameras and five wearable sensors.

Figure 5.6: Real-world implementation of our proposed approach. Our working setup in Figure 5.5 is shown here recording audio-visual events for evaluating crossmodal synchronization, as discussed in Section 5.4.2.

we discuss details in Section 5.5. 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 5.3.3.

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 [239]. 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.

5.3.2 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 [240].
- 48 custom wearable sensors adapted from the open source Rhythm Badges [63]; 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 5.6. 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 5.4. 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 5.3).

Relaying time to cameras. We explain the bottom branch in Figure 5.3 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) [241]. The LTC signal that is converted from NTP is sent to a base station unit by Timecode Systems called :pulse [242], 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 [243], also manufactured by Timecode Systems. Each Syncbac PRO is physically tethered to a GoPro camera so 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) [242, 243].

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 5.4 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 5.4. 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 5.3 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 [63] 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 [244]. 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 [244]. 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.

5.3.3 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 [245], with a specific example finding a 157 ms latency in speech perception [246]. 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 [216]. Studies in synchrony, mimicry, entrainment, and other higher-level social phenomena usually consider a larger window size. Levitan et al. [217] 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 [67, 247].

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 [248]. 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 μ s and wearable data latency of 5 ms over 9.5 hours [244]—is also applicable to data collection situations where fine details like faces are important such as egocentric vision setups, or those involving physiological sensors.

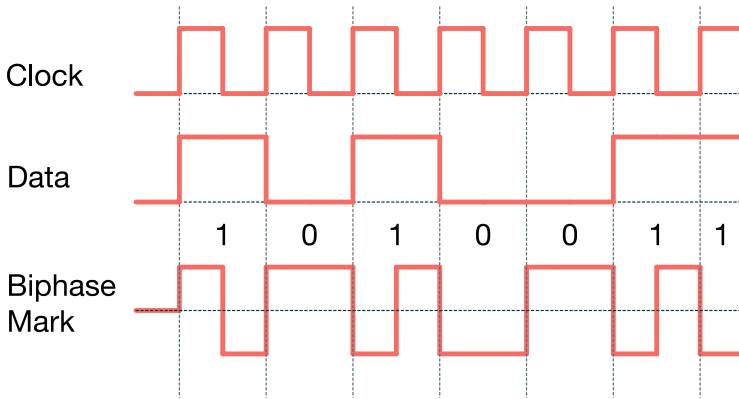


Figure 5.7: Biphase Mark Encoding of Linear Time Code

5.4 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 [45, 227]. As discussed in Section 5.3.2, 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 5.4.2. 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 5.4.1. Latency measures in our individual sensor subnetworks are depicted in Figure 5.3 and already discussed in Section 5.3.2.

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.

5.4.1 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

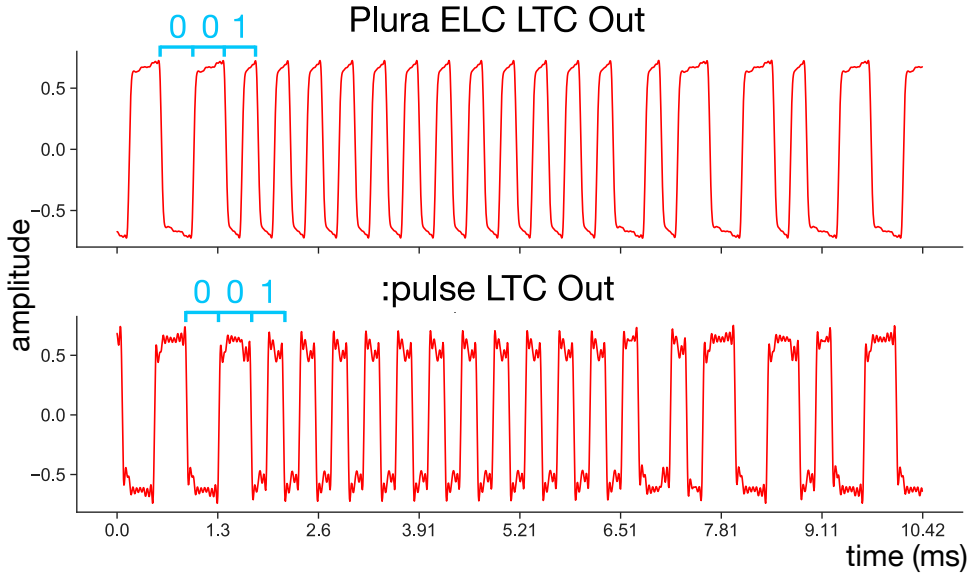


Figure 5.8: Raw audio LTC signals generated by the Plura ELC and :pulse modules. The window includes the encoding of an LTC sync word (00111111111101) followed by the bits 0001000 from the next frame. The lower signal here leads the upper signal by 62 audio samples, or less than 1 bit of data.

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 5.7: 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’ 00111111111101 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 [249] through custom cables. Figure 5.9 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 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 5.8 depicts a window from our

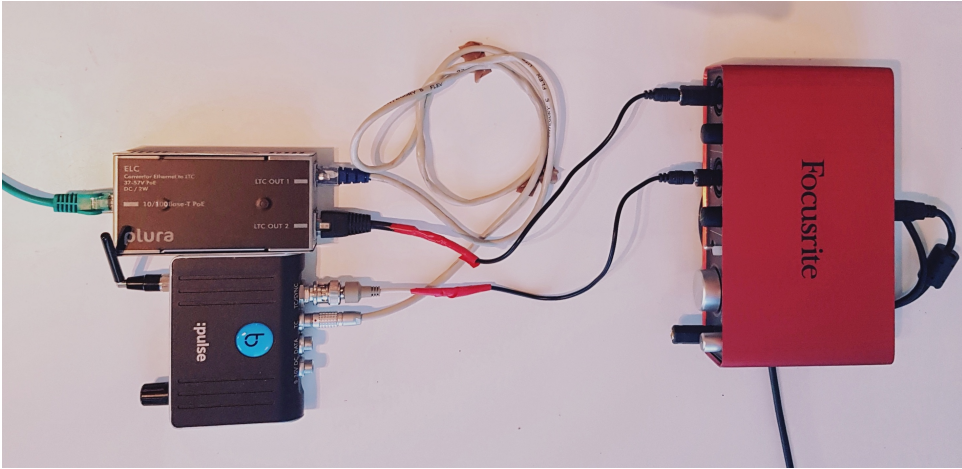


Figure 5.9: Hardware setup with custom cables for recording LTC signals from the Plura ELC and the :pulse base station.

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 5.7.

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 5.3.3.

5.4.2 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 5.5 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 5.10 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 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 [244, 250]. 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 [244]. 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 5.5) 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 5.3.3 and the human perception tolerance limit of audiovisual skew which is ± 80 ms [251]. 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.

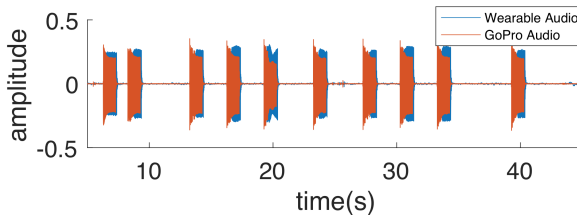


Figure 5.10: Representative example showing the aligned audio events in one of the wearable sensors and one of the GoPros.

5.5 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 [252].

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 5.3.1. Additionally, GPS modules can be used to generate NTP and PTP signals [253] 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 [254]. 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.

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

5.6 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 superseded 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 $414\ \mu\text{s}$ 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 [244]. 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. 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.

6

CONFLAB: A DATA COLLECTION CONCEPT, DATASET, AND BENCHMARK FOR MACHINE ANALYSIS OF FREE-STANDING SOCIAL INTERACTIONS IN THE WILD

6

This chapter is accepted for publication as: Chirag Raman*, Jose Vargas-Quiros*, **Stephanie Tan***, Ashraful Islam, Ekin Gedik, Hayley Hung. Conflab: A Rich Multimodal Multisensor Dataset of Free-Standing Social Interactions In-the-Wild. **Conflab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild**. Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS) Dataset and Benchmark track, December 2022.

*: equal contribution



Figure 6.1: 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.

6.1 INTRODUCTION

A crucial challenge towards developing artificial socially intelligent systems is understanding how *real-life* situational contexts affect social human behavior [255]. Social-science findings indeed show that the dynamics of how we conduct daily interactions vary significantly depending on the social situation [256–258]. Unfortunately, such dynamics are not adequately captured by many data collection setups where role-played or scripted scenarios are typical [259].

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 6.1).

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 [13, 260, 261]. The most common solution involves mounting cameras from aerial perspectives such as top-down [262] and elevated-side views [45, 46, 263]. 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 [264]. Besides, state-of-the-art (SOTA) body-keypoint estimation techniques perform poorly on aerial perspectives [172, 262], making the extraction of automatic pose annotations challenging (Figure 6.3). To avoid such issues, some researchers have turned to more privacy-preserving wearable sensors shown to benefit many behavior analysis tasks [13, 24, 121].

In all, the closest related datasets (see Figure 6.4) 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 [265, 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.

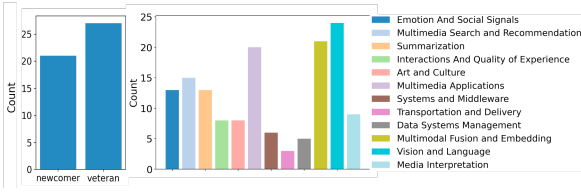


Figure 6.2: Frequency of newcomer/veteran participants (left) and reported research interests (right).

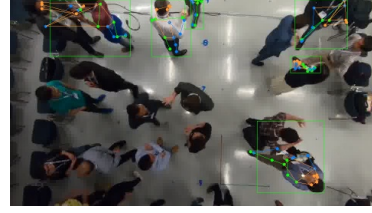


Figure 6.3: Keypoint detection using pre-trained RSN [266]. Additional SOTA results are in Appendix 7.5

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 6.2). This was achieved by organizing ConfLab as part of a major international scientific conference. ConfLab had these goals: (i) a data collection effort following 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 [264, 267], 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 recognition of social actions (under-explored in the top-down perspective), (c) pose-based F-formation estimation (has not been possible from prior work [6, 17, 60, 268]), and (d) the direct study of interaction dynamics using full body poses (previously limited to lab settings [269]). **(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 [121]. **(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.

Dataset	People/ Scene	Video	Manual Annotations	Wearable Signals	Synchronization
Cocktail [50] †	7	512 × 384	F-formations (20 and 30 min, 1/5 Hz)	None	Unknown
CoffeeBreak [22]	14	1440 × 1080	F-formations (130 frames in two sequences)	None	None
IDIAP [19]	> 50	180 min; 654 × 439 20 fps	F-formations (82 independent frames)	None	None
SALSA [45] †	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[41] †	32	30 min; 1920 × 1080 30 fps	Bounding boxes (30 min, 1 Hz ‡) Head & body ori. (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 [25] BT: Bluetooth IMU: Inertial Measurement Unit					

Figure 6.4: 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.

6

6.2 RELATED WORK

Early datasets of in-the-wild social events either spanned only a few minutes (e.g. Coffee Break [263]), 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 [17]). More recently, two different strategies have emerged to circumvent such issues.

One approach involves fully instrumented labs with a high resolution multi-camera setup for video and audio data. Here automatic detectors [214, 269, 270] 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 [43], and more recently standing scenarios like the Panoptic Dataset [269]. 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 [45, 262]. While proximity has been used as a proxy of face-to-face interaction [45, 223, 271–273], recent findings highlight significant problems with such an assumption [274]. 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 Cabrera-Quiros et al. [262] show more promising results with a radio-based proximity sensor and ac-

celerometer [275], but such data remains insufficient for more downstream tasks due to the relatively low sample (20Hz) and annotation (1Hz) frequency [121]. 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 [45] and MatchNMingle [262] 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 [121, 276, 277]. While 1 Hz is able to capture some conversation dynamics [160], it is insufficient to study fine-grained social phenomena such as back-channeling or mimicry that involve far lower latencies [265, 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 [265] (see Sec. 6.3). Figure 6.4 summarizes the differences between Conflab and other related datasets.

6.3 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 [13, 260, 261]. 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 6.5). 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 (TU Delft) and the conference location's national authorities (France). All participants gave consent for the recording and sharing of their data at registration. (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 [278, 279].

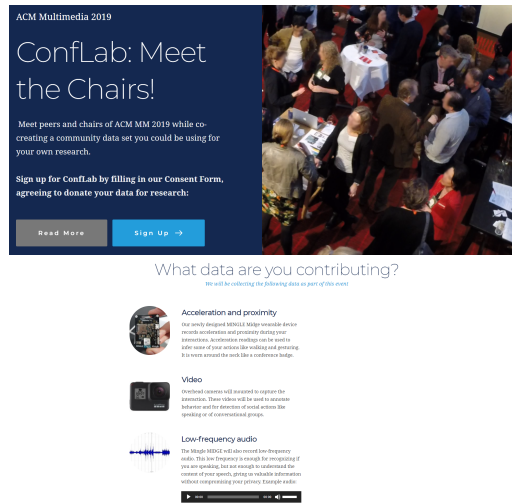


Figure 6.6: The Midge

Figure 6.5: Screenshots from the *ConfLab: Meet the Chairs!* event website

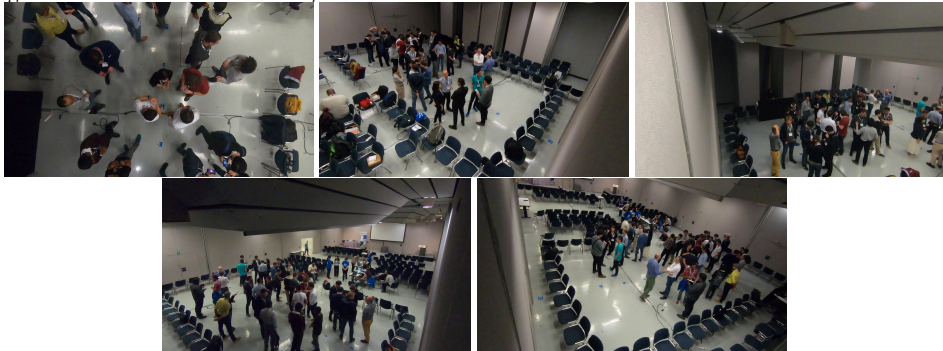


Figure 6.7: Comparing the top-down (top-left, camera 4) and elevated-side camera views (rest). Note how the top-down view is better at mitigating the capture of faces and suffers from fewer occlusions. This allows for a clearer capture of gestures and lower extremities for the most number of people while also preserving privacy.

Data 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 [264]. 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 6.7). The 10 m × 5 m interaction area was recorded by 14 GoPro Hero 7 Black cameras (60fps, 1080p, Linear, NTSC) [240]. 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 (see Figure 6.6 for a design render), based on the open-source Rhythm Badge designed for office environments [69]. We improved upon the Rhythm Badge to achieve more fine-grained and flexible data capture (see Appendix 7.5). 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 [45], or speech plosives [227]. 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 [265], making it easier for others to replicate our capture setup. See Appendix 7.5 for synchronization and calibration details, and Appendix 7.5 for images of the setup.

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 [280, 281]. 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 [282] 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.

6.4 DATA ANNOTATION

Continuous Keypoints Annotation Existing datasets of in-the-wild social interactions have mainly focused on localizing subjects via bounding boxes [45, 262]. 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 a multiple skeletal keypoints. The typical approach to keypoint annotation involves using tools such as Vatic

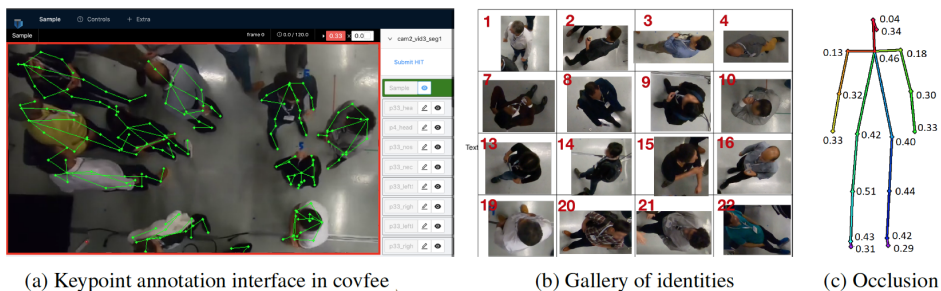


Figure 6.8: Illustration of the body keypoints annotation procedure: (a): our custom time continuous annotation interface; (b): the gallery of person identities used by annotators to identify people in the scene (faces blurred); and (c): the skeleton template with the fraction of occluded frames.

[283] or CVAT [284] 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 [285]. 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 [284] is presented in [285]. For the pilot annotators, the continuous annotation methodology resulted in a $3\times$ 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 6.7). Given the overlap in the camera views, we annotated keypoints in five of the ten overhead cameras (see Figure 6.1). 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 [286]. We annotated the binary speaking status of every subject due to its importance as a key feature of social interaction [24, 35, 81, 287, 288] and to contribute the existing community who are working on this task [121, 289, 290].

Action annotations have traditionally been carried out using frame-wise techniques [262], 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 [285]. 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 [276]. Similar to [262], the annotations were 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 [45, 46, 262], F-formations group membership were annotated using an approximation of Kendon's definition [23]. 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.

6.5 DATASET STATISTICS

Individual-Level Statistics Figure 6.8(c) (shows the average occlusion values we obtained from annotators for each of the 17 keypoints. In Figure 6.9(a) 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.

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 fig:annostats(b) and Figure fig:annostats(c), 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 fig:annostats(d). 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.

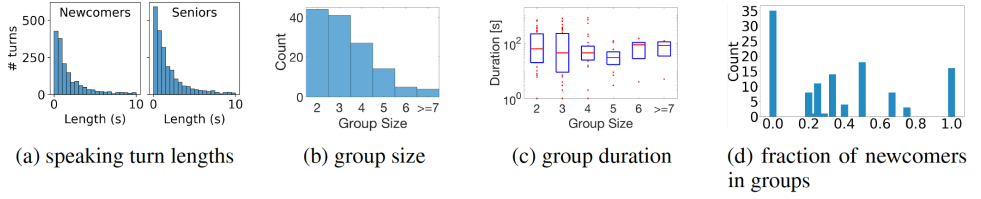


Figure 6.9: Data distributions for speaking status and conversation groups

6.6 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.

6

6.6.1 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 [172] (also see Figure 6.3 and Appendix 7.5 for Conflab results), we finetune COCO-pretrained models on our dataset. We used Mask-RCNN [291] (Detectron2 framework [292] 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 7.5.

Evaluation Metrics We evaluated person-detection performance using the standard metrics in the MS-COCO dataset paper [293]. 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) [?]. AP^{OKS} is a mean average precision for different OKS thresholds from 0.5 to 0.95.

Results and Analyses Table 6.1 summarizes our person detection and joint estimation results. Our baseline achieves 73.9 AP_{50} in detection and 45.3 AP_{50}^{OKS} in keypoint estimation. Figure 6.10 shows qualitative results from our fine-tuned network. For further insight we performed several analyses and ablations. In Appendix Table 2, 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

Table 6.1: Mask-RCNN results for person bounding box detection and keypoint estimation.

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

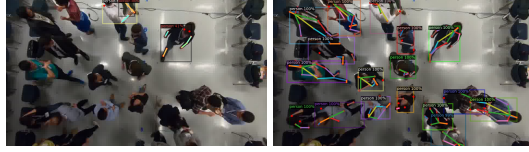


Figure 6.10: Predictions from the Mask-RCNN model; COCO pretrained (left), and ConfLab finetuned (right).

cameras one at a time. We report results in Appendix Table 3. 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 4.

6.6.2 SPEAKING STATUS DETECTION

In data collected from real-life social settings, individual audio recordings can be hard to obtain due to privacy concerns [294]. This has led to the exploration of other modalities to capture some of the motion characteristics of speaking-related gestures [276, 277]. 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 [295], 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 [296], InceptionTime [297], and Minirocket [298]. We performed late fusion by averaging the scores from both modalities. Like prior work [121, 277], 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 [295]. 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 6.2 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 [121]. Multimodal results were slightly higher than acceleration-only results, despite our naive

Table 6.2: 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 [300]	0.676	0.677
	InceptionTime [297]	0.798	0.768
Acceleration	Resnet 1D [296]	0.801	0.767
	Minirocket [298]	0.813	0.768
Multimodal	MS-G3D + Minirocket	0.823	0.775

Table 6.3: Average F1 scores for F-formation detection comparing GTCG [268] and GCFF [129] 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)

fusion approach, a possible point to improve in future work [299]. Experiments with the rest of the IMU modalities are presented in Appendix 7.5.

6.6.3 F-FORMATION DETECTION

Setup Like prior work [6, 17, 60, 268], we operationalize interaction groups using the framework of F-formations [23]. We provide performance results for F-formation detection using GTCG [268] and GCFF [129] as a baseline. Recent deep learning methods such as DANTE [60] 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 [46]) and tuned a subset of parameters more relevant to Conflab attributes on camera 6. More details can be found in Appendix 7.5. 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 [129, 268]. 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 6.3.

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 [23]) may have multiple conversations floors [81]. Floors are indicated by coordinated speaker turn taking patterns and influence coordinated head orientations of the group.

6.7 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 [17, 262] 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 [14], dominance [288], rapport [35], influence [301] 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 [302]. 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 [303]. 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 uniquely dependent not only on the individual, but also the group they are interacting with [304]. We hope our highlighting of participatory design practices and these value-sensitive design principles promote social safety in developing socially assistive technologies.

7

DISCUSSIONS AND CONCLUSION

The common theme in this thesis is to use multimodal data including vision, audio, and wearable sensing to estimate human head and body orientations which are important cues in social interactions, aiming to capture both temporal and social dynamics. The approaches were validated by data captured in-the-wild real life scenarios, presented by modeling-oriented works in Chapter 2, 3, and 4, and sensor-related and data-related works in Chapter 5 and 6. Below I first summarize learnings from each chapter and follow with a general discussion of open topics related to this thesis.

Chapter 2 presents an approach to address the limitation in head orientations in in-the-wild setting. The proposed solution uses a combination of proxemics (positions and orientations) and dynamics (speaking status, body motion through accelerometer) features as temporal inputs to further capture the intrinsic temporal dynamics of the signals. The deep learning architecture also incorporates a context pooling layer that models the coupled head behaviors between members of the same conversation group. We showed the advantage of using this feature set and the proposed model to achieve better performance compared to baselines that do not model the temporal or group dynamics. A thorough comparison between the different feature sets was also included, though unsurprisingly, body orientations are still the most important feature towards estimating head orientations. More importantly, there is significant performance improvement over time segments where turn-taking occurs. However, the performance of the trained model for head orientation estimation is largely dependent on the quality of the annotated ground truth, which is shown to be challenging using the top-down privacy preserving camera view. We showed that there is a slight improvement in using a regression versus a classification setup, as we aim to capture head turns which could be indicative of high-level social cognitive phenomena and frames possible research direction.

Chapter 3 presents an approach to F-formations detection with spatio-temporal context. The approach uses temporal sequences of behavioral cues such as positions, head, and/or body orientations, and includes a spatial context pooling module that accounts from the surroundings when predicting pairwise affinities. Our deep learning architecture design allows for estimation of affinity jointly between the current person and everyone else in the scene. This was an important design choice to allow for different group sizes

whilst accounting for group context. We showed that with temporal context, our method outperforms baseline methods [6, 53, 60] on the Conflab dataset which has higher temporal resolution compared to other datasets [46, 68]. Instead of focusing only on F-formation detection evaluation after acquiring pairwise affinity scores like in prior works, we show an analysis of how the accuracy of affinity scores relate to group detection. We also demonstrate how the affinity scores change over time, which could be indicative of future group memberships. A comparison between different preprocessing of affinity scores before applying Dominant Set clustering for group extraction was shown. We found that averaging the pairwise affinity scores leads to a better performance compared to taking the raw predicted values or taking the maximum or minimum of the two pairwise affinity scores. An explanation could be the binary ground truth for affinity values is symmetric, which is implicitly captured by the model. A comparison for performance on scenes according to dynamic scene change events also shows that the proposed method outperforms the baseline method on scenes that have more group (re)formations and disbanding. Additionally, to showcase the usability of the predicted affinity scores, a forecasting framework was proposed to predict future conversation groups with uncertainty quantification. Future works related to this chapter could include utilizing more high frequency inputs to capture sub-second social dynamics, such as turn-taking and body movement synchrony and mimicry.

In Chapter 4, we present a solution for estimating head and body orientation using wearable sensors when there is a limited amount of annotations. Because of the multi-camera view with heavy occlusion, annotations of head and body orientations can be expensive to obtain from video cameras. This approach takes advantage of all available modalities in the dataset (audio and proximity, in addition to video) to obtain weak labels of head and body orientations. It also accounts for the temporal smoothness that exists in head and body orientations with an improved module using Gaussian Process Regression, as well as the anatomical constraints in head and body coupling. Similar to [28, 29] in using the matrix-completion approach under the transductive setting, we formulated the task into an optimization problem that flexibly models constraints, while accommodating varying amount of missing data (in head and body weak labels). Using only 5% of the manually annotated labels, this approach enables 68% and 76% classification accuracy in head and body orientation estimation, respectively. The paper contains an extended analysis with experimental results of the contribution of each of the module in the optimization problem, and showcases the value of using head and body weak labels in this task. This approach combines considerations of data scarcity and the availability of multimodal data in estimating head and body orientations, and could serve as a solution to propagate labels in datasets.

In Chapter 5, we address the limitation in multi-modal data synchronization which are traditional manual or event-based. One of the major challenges in developing methods to model human social behavior is the lack of datasets in-the-wild. Existing data collection methods have mostly been situated in lab-based settings which are difficult to scale up to capturing social interactions that span a large physical space such as a conference venue. The post-processing of different streams of data collected in unscripted scenarios in-the-wild has relied on event-based and manual synchronization such as balloon pops, clapboards, etc. Instead, our solution offers hardware-based temporal alignment of the

signals which is more precise and could be important to down-stream modeling of these signals. Our proposed solution enables a common timestamp based on Network Time Protocol (NTP) in both wearable sensors and camera network. It is a cost-effective, modular, reproducible, and scalable open-source approach that unifies NTP timestamps which offers cross-modality synchronization between video, audio, and wearable sensing signals within 13 ms. We validated this approach by quantifying the latency introduced by the components that convert NTP to Linear Time Code (LTC) signals used by the video recordings. We also quantified the audiovisual latency in the events captured by wearable sensors and video cameras. The maximum delay was found to be within 13 ms which suffices studies in modeling human social behaviors such as synchrony in body movement on the scale of 40 ms. A limitation that we discussed is related to further improvement in synchronization accuracy which could be offered using GPS time signals, as well as the scalability of Bluetooth hubs and video camera RF base stations, as there could be additional delays introduced by those networks.

To address the lack of in-the-wild social interaction dataset, Chapter 6 presents a new dataset, Conflab. It contains high-fidelity and fine-granularity measurements and annotations of social interactions at a professional networking event at ACM MM 2019. Using the data synchronization and acquisition method presented in Chapter 5, the multimodal data are temporally aligned which allows for downstream social behavior modeling. The Conflab dataset contains skeletal keypoint annotations of 48 individuals and their speaking status at 60 Hz annotation frequency, conversation group membership at 1 Hz, and meta-data related to interests and seniority of group members. With illustrated benchmarks of automated person and keypoint detection, multimodal speaking status detection, and conversation group membership estimation using a variety of orientation cues, Conflab provides new opportunities for research in automated methods that explore full-body synchrony estimation, full-body F-formation, in relations to speech dynamics and turn-taking patterns.

While each of these chapters addresses their open questions respectively, some related but unexplored topics which require further investigation are discussed in the following sections.

7.1 ACCURATE PERCEPTION OF HUMAN INTERACTIONS IN-THE-WILD

This thesis has focused on the estimation of head and body orientations in crowded social scenarios (Chapter 2 and 3) but has not addressed their connection to visual focus of attention (VFOA) (i.e., where or who people are looking at). VFOA is an important social cue for social attention that could be indicative of the quality of interaction [176]. A better understanding of the dynamics of VFOA enables better implementations of machines with natural human-like interaction abilities. However, in the elevated side-view or overhead views (mostly used in this thesis for privacy concerns and to preserve ecological validity), it is unclear how to obtain an accurate measure of VFOA, as the limitations in head orientation estimations are partially due to annotation noise. Eye gaze information, which is normally more representative of the VFOA, is still absent and would otherwise be hard to capture without disturbing the ecological validity of the social interactions.

A potential answer to this challenge lies in adopting an egocentric viewpoint as an additional perspective to the surveillance point of view. Recent technology such as the Tobii and Pupil Invisible eyeglasses [305, 306] are minimally invasive eyeglasses that track eye and head movement with a design focus of allowing studies of human behaviors. Combining the first-person (egocentric) and third-person (allocentric) view allows us to better understand the connection between individual behaviors and group behaviors and other emergent social processes. It also addresses the challenges in measuring human behaviors and obtaining reliable annotations introduced in the Chapter 1. Because these egocentric are designed to be minimally invasive, it preserves the ecological validity to a large extent, which is important in modeling human behaviors. This combination of different perspectives has already been proven to be effective in person re-identification [307], as well as joint attention learning [308].

However, there is very limited resource to relate third person videos and first person videos which are applicable to studying natural human social interactions in-the-wild. Leveraging the existing experience in using third-person visual and wearable modalities (e.g., this thesis), a hybrid approach of first-person video, third-person video and wearable sensing data, enables capturing higher fidelity human interaction data without disrupting the ecological validity of interactions. This in turn will allow more accurate ground truth information such as VFOA. Subsequently, a connection may be drawn between VFOA and cognitive attention, and thereby enabling probing into conversational quality and roles. This is to partially address the annotation noise in head orientation estimation (18 degrees) as provided in Chapter 2. Other interesting annotations include facial action units and expressions of interactants, turn-taking, events such as (joint) laughter, conversation floors, experience sampling measures of interaction quality, and various phenomena which are difficult to measure without egocentric perception. Ultimately, these annotations could provide better measurements of interaction quality (i.e., conversation quality [14]) compared to head and body orientations

Apart from the benefits that a hybrid approach brings, standing technical challenges include mapping egocentric gaze to 3D human targets or objects in the scene from the wearable eye tracker. This could be achieved by 3D reconstruction with methods that estimate the wearer's gaze depth (e.g., [309]). Secondly, synchronizing eyeglasses data with other modalities requires significant integration with our existing framework. We can synchronize these audio-visual-wearable data via network time protocol or event-based methods. Although commercial glasses are wireless, integrating them with this synchronizing framework on a hardware level still requires considerable amount of work. These design choices during data collection are important and intertwined with the outlooks of answering research questions.

7.2 THE SCALE OF MODELING HUMAN BEHAVIORS

In this thesis (Chapter 2-6), the scale at which human behaviors and interactions are modeled is at the level of complex conversation scenes, which increases the difficulty of head/body orientation estimation, compared to meeting-like scenarios (i.e., seated and more structured scenarios). The choice of automated methods for human behavior modeling is dependent on the scale and context at which these behaviors are observed. In this section, I discuss the choice of using visual, wearable sensing, and combination thereof for measuring

and modeling human behaviors.

In the decade prior to this thesis, computer vision-centric works using visual features have already investigated tasks including person detection (e.g., [310–312]), tracking [313, 314], head/body pose estimation [117, 315]. These methods achieve accurate predictions of orientations in all three degrees of movement (roll, pitch, and yaw) in constrained settings such as meeting or for pedestrians, but are usually not good enough for downstream more socially-relevant tasks, such as visual attention estimation or conversation group detection, for free-standing conversation groups. Challenges include occlusion and performance reliability when the scene involves between 30 and 50 subjects. The estimation for all the subjects need to reach a level of accuracy where the downstream tasks such as conversation group membership can be applied (if the method rely on orientation features).

While this thesis investigates the multimodal aspect of improving the performance in these tasks, the visual modality is still the mainstream modality of sensing, and especially suitable for orientation estimation. To improve over deep-learning based visual features, recent methods in the computer vision community could be exploited to improve orientation estimation at this scale. For example, 3D skeletal reconstruction of humans using monocular cameras is a promising direction. This class of approaches relies on the abundance of visual information, techniques, and modeling expertise from the vision community, and more importantly, is a super set of the orientation estimation task at hand. If 3D skeletal reconstruction can be reliably achieved, orientation information will be automatically available. The reconstructed 3D skeletal information allows new research questions that explore body motion from a visual perspective, which could better detect phenomena such as synchrony and mimicry. With sufficient surveillance cameras (visual information), orientation estimation for a large scene used in an application use-case (e.g., social surveillance) could be a tangible goal. Since the visual modality is relatively accessible (through surveillance), the vision based methods are also scalable.

The combination of vision and wearable sensing modalities (or their derived features) for estimating human behaviors (Chapter 2, 3, and 4) finds a potential use case in professional networking events. The datasets used in this thesis fit in this setting, where multimodal measurement and perception is possible. However, there is a trade-off to be considered: the value of wearable sensing adds to automatic orientation estimation vs. the overhead of developing and deploying wearable sensors to different events which can be different for every instance. If multimodal methods indeed prove to be exceptionally advantageous, resources put into multimodal sensors and approaches are better justified. Consequently, the sensing capability needs to be further assessed, and the methods to join these sensor streams need to be further developed, for scalability.

On the other extreme of the size scale (where the setting spans a large physical space and when it does not make sense to install many surveillance cameras), wearable sensing is an alternative way to estimate human interactions because of its ubiquitous nature. The details of using wearable sensors for studying human interactions are discussed in Sec. 7.3.

In summary, the choice of using cameras, wearable sensors, or combination depends on the size of the scene, accessibility of sensors, the granularity of human behaviors one wishes to capture, and ultimately the application in question.

7.3 CONCERNS FOR WEARABLE SENSORS

In the Social Signal Processing community, many wearable sensors have been developed to study social interactions, including the popular Sociometric Badges (Sociometers) [221]. Recent sensors improve upon previous ones in terms of capability and inclusion of more sensing modalities. For intended use, Sociometers lay out clear goals: to measure the amount of face-to-face interaction, conversational dynamics (e.g. turn taking patterns, tone of voice, etc.), proximity to other people, and motion activity level using social signals derived from vocal features, body motion, and relative location. However, in both Sociometers and the more recent Rhythm Badges, the design choice was made to downsample the audio recordings. It has also been found that the timestamps of the signal recorded by the Sociometers could deviate from true time by more than 1 minute [316]. Issues related to the inflexibility in working with Sociometers, as well as the reported time synchronization misalignment motivates future research and sensor development.

Even though the Sociometers have been commercialized, many problems were only discovered after extensive testing and usage (such as the need for raw data for researchers). These problems are often unaddressed because the people who are the most knowledgeable about the sensors are temporary contracted (most likely PhD and Postdoc researchers) [274]. New requirements also arise over time and because of the lack of systematic standards and approaches, they remain as ad-hoc modifications implemented by only the ones in need. With the discontinuation of the company, support with respect to the sensors has also reduced. The follow-up development of Sociometers, the Rhythm Badges, was open-sourced, and also already alleviated some of the problems related to inflexibility and need for customization. While the lead developers have established communities for exchange, users (most often researchers from other groups) largely remain on their own for clarifications and/or extensions of the sensors. There is no existing documentation of the performance (similar to a specification sheet for hardware), as well as the known and potential limitations of these custom built sensors. Compared to open source software code, hardware projects typically need more specification and assessment for reproducibility.

As another example for custom wearable sensors, the Midges (developed for Chapter 6) are built to address specific research questions in social signal processing. While the Midges are open-sourced and scalable for manufacturing and production by design, deployment, peripheral infrastructure, and firmware need to be tested and developed more extensively in order to guarantee performance and future improvement over the sensors. Iteration of design (if any) and active user support are also part of the process. For example, suppose the use case is to expand for the ubiquitous computing community where scalability is a requirement, the corresponding documentation needs to be availability and perhaps as well as a responsible point-of-contact. When there is a need for widespread adoption of the wearable sensors in more real-life in-the-wild settings, practices in productizing the sensors more formally may be considered.

7.4 REPRODUCIBLE DATA COLLECTION

This thesis, specifically Chapter 5 and 6, revolves around easing the data collection process from human social interaction studies and specifically for complex conversational scenes in-the-wild. Reproducibility, in terms of results replication can be achieved through open

source code sharing, and is getting appropriate attention from the community. However, not enough focus is placed on the reproducibility of data collection experiments. For social signal processing studies of in-the-wild human social behaviors, data collection is one of the most challenging aspects of the research. There is considerable ground work that could be easily dismissed but actually consists of important design choices and considerations: what events to capture, how many subjects, composition of the subjects, the venue of the event, etc. After the data is collected, annotators need to be recruited and trained for labelling the data. In some cases, experts (e.g. psychologists) need to be involved to decide the correct labelling of the interaction phenomena in question. These pieces are intricately connected and difficult to modularize. With these considerations, there is an increasing need to systematize data collection in this field, by making a template, checklist for documentation, etc., with each aspect appropriately catered for different modalities. Even though each instance of data collection is generally ad-hoc, research groups that collect dataset often should share their information as it could be beneficial to other practitioners. Additionally, by attempting to establish a more standard operating procedure for dataset collection, users of the data can be better educated and also become aware of how the data is collected to a certain extent, as the methods they develop based on the data may ultimately impact the applications.

7.5 MODELING HUMAN INTERACTIONS: A DIFFERENT VIEWPOINT

In the methods-oriented work in this thesis (Chapter 2, 3, and 4), I focused on developing automated methods to model human head and body behaviors in complex conversational scenes, where the interpersonal dynamics are specific to the setting. While this data-driven approach is justified, the underlying task, which is to understand and model human interaction and coordination patterns, is also studied by psychologists, physicists, etc., using their respective approaches. A particularly interesting angle to view human interaction behaviors as a complex dynamical system. The key question is to ask whether the social processes occur on multiple scales adhere to general principles, a form of universal laws.

Existing frameworks of modeling coordination typically focus on either very large number of elements or systems with few elements. In the first case, most models are based on statistical mechanics. In the latter case, the synergetics and nonlinear dynamics are modeled by the Haken-Kelso-Bunz (HKB) model (involving 2 or 3 entities) [317]. HKB has been adapted and extended to model increasingly complex scenarios, such as for many diachronic social behaviors like turn-taking in conversations [318]. While the laws of statistical mechanics may be appropriate in studying a large crowd of people and not applicable to analyzing complex conversational scenes, HKB could be adapted to model human coordination in scenes such as cocktail parties and networking events. However, modifications to the HKB model, including scaling up for extrapolation, are challenging and still unresolved. An interesting direction to explore is the fusion of this model-based approach and the data-driven approach (such as this thesis) to discover a more general framework of how humans interact socially.

Another approach is agent-based modeling, which involves simulations of artificial agents interacting over time within specific context or environment [319]. In this case,

agents represent humans who have pre-specified attributes and behave in specifically defined ways. Even though the agent behaviors are simplified, the outcome is typically not trivial, as local interactions between agents give rise to large-scale dynamics on the group- and system- level. This is particularly apt for discovering emergent behaviors, where the combination of small-scale individual behavior creates different collective behavior. This approach may allow more focus on understanding the dynamic system as a whole. Once more understanding is achieved, the outcome of a model is not limited to a set of predictions but also a set of steps or strategies to mitigate unwanted effects in a dynamic system, which ultimately may be more useful from an application point of view.

BIBLIOGRAPHY

REFERENCES

- [1] Alan Page Fiske. The cultural relativity of selfish individualism: anthropological evidence that humans are inherently sociable. 1991.
- [2] Michael Argyle. *Social interaction*. Routledge, 2017.
- [3] Douglas T Kenrick, Steven L Neuberg, and Robert B Cialdini. *Social psychology: Unraveling the mystery*. Pearson Education New Zealand, 2005.
- [4] Derek Freeman. Social anthropology and the scientific study of human behaviour. *Man*, 1(3):330–342, 1966.
- [5] Michael Argyle, Adrian Furnham, Jean Ann Graham, et al. *Social situations*. Cambridge University Press, 1981.
- [6] Francesco Setti, Chris Russell, Chiara Bassetti, and Marco Cristani. F-formation detection: Individuating free-standing conversational groups in images. *PloS one*, 10(5):e0123783, 2015.
- [7] Norbert L Kerr and R Scott Tindale. *Methods of small group research*. 2014.
- [8] Erving Goffman. The interaction order: American sociological association, 1982 presidential address. *American sociological review*, 48(1):1–17, 1983.
- [9] Adam Kendon. Goffman’s approach to face-to-face interaction. *Erving Goffman: Exploring the interaction order*, 1988.
- [10] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.
- [11] Edward T Hall, Ray L Birdwhistell, Bernhard Bock, Paul Bohannon, A Richard Diebold Jr, Marshall Durbin, Munro S Edmonson, JL Fischer, Dell Hymes, Solon T Kimball, et al. Proxemics [and comments and replies]. *Current anthropology*, 9(2/3): 83–108, 1968.
- [12] Adam Kendon. Movement coordination in social interaction: Some examples described. *Acta psychologica*, 32:101–125, 1970.
- [13] Hayley Hung, Ekin Gedik, and Laura Cabrera Quiros. Complex conversational scene analysis using wearable sensors. In *Multimodal Behavior Analysis in the Wild*, pages 225–245. Elsevier, 2019.

- [14] Chirag Raman, Navin Raj Prabhu, and Hayley Hung. Perceived conversation quality in spontaneous interactions. *arXiv preprint arXiv:2207.05791*, 2022.
- [15] Chirag Raman, Hayley Hung, and Marco Loog. Social processes: Self-supervised forecasting of nonverbal cues in social conversations. *arXiv preprint arXiv:2107.13576*, 2021.
- [16] Stephen RH Langton, Roger J Watt, and Vicki Bruce. Do the eyes have it? cues to the direction of social attention. *Trends in cognitive sciences*, 4(2):50–59, 2000.
- [17] Hayley Hung and Ben Kröse. Detecting f-formations as dominant sets. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 231–238, 2011.
- [18] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):190–204, 2017.
- [19] Sileye O Ba and Jean-Marc 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, 2009.
- [20] Benoît Massé, Silève O. Ba, and Radu Horaud. Tracking gaze and visual focus of attention of people involved in social interaction. *CoRR*, abs/1703.04727, 2017.
- [21] Kazuhiro Otsuka, Yoshinao Takemae, and Junji Yamato. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 191–198, 2005.
- [22] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. Social interaction discovery by statistical analysis of f-formations. In *BMVC*, volume 2, page 4, 2011.
- [23] Adam Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.
- [24] Ekin Gedik and Hayley 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):163, 2018.
- [25] Kaitlin EW Laidlaw, Tom Foulsham, Gustav Kuhn, and Alan Kingstone. Potential social interactions are important to social attention. *Proceedings of the National Academy of Sciences*, 108(14):5548–5553, 2011.
- [26] Kim Kopenhaver Haidet, Judith Tate, Dana Divirgilio-Thomas, Ann Kolanowski, and Mary Beth Happ. Methods to improve reliability of video-recorded behavioral data. *Research in nursing & health*, 32(4):465–474, 2009.

- [27] Ipke Wachsmuth, Manuela Lenzen, and Günther Knoblich. *Embodied communication in humans and machines*. OUP Oxford, 2008.
- [28] Xavier Alameda-Pineda, Yan Yan, Elisa Ricci, Oswald Lanz, and Nicu Sebe. Analyzing free-standing conversational groups: A multimodal approach. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 5–14, 2015.
- [29] Ramanathan Subramanian, Yan Yan, Jacopo Staiano, Oswald Lanz, and Nicu Sebe. On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction, ICMI '13*, pages 3–10, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2129-7. doi: 10.1145/2522848.2522862. URL <http://doi.acm.org/10.1145/2522848.2522862>.
- [30] Crystal L Hoyt, George R Goethals, and Ronald E Riggio. Leader-follower relations: Group dynamics and the role of leadership. *The quest for a general theory of leadership*, 1(1):96–122, 2006.
- [31] Dinesh Babu Jayagopi, Hayley Hung, Chuohao Yeo, and Daniel Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):501–513, 2009.
- [32] Hayley Hung, Yan Huang, Gerald Friedland, and Daniel Gatica-Perez. Estimating dominance in multi-party meetings using speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):847–860, 2010.
- [33] Hayley Hung and Gokul Chittaranjan. The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 879–882, 2010.
- [34] Gokul Chittaranjan and Hayley Hung. Are you awerewolf? detecting deceptive roles and outcomes in a conversational role-playing game. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5334–5337. IEEE, 2010.
- [35] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In *23rd International Conference on Intelligent User Interfaces*, pages 153–164, 2018.
- [36] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3477–3485, 2019.
- [37] Bernd Dudzik, Michel-Pierre Jansen, Franziska Burger, Frank Kaptein, Joost Broekens, Dirk KJ Heylen, Hayley Hung, Mark A Neerincx, and Khiet P Truong. Context in human emotion perception for automatic affect detection: A survey of audiovisual databases. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 206–212. IEEE, 2019.

- [38] Kazuhiro Otsuka, Junji Yamato, Yoshinao Takemae, and Hiroshi Murase. Conversation scene analysis with dynamic bayesian network based on visual head tracking. In *2006 IEEE International Conference on Multimedia and Expo*, pages 949–952. IEEE, 2006.
- [39] Laurence A Clarfeld, Robert Gramling, Donna M Rizzo, and Margaret J Eppstein. A general model of conversational dynamics and an example application in serious illness communication. *Plos one*, 16(7):e0253124, 2021.
- [40] RC Schmidt, Samantha Morr, Paula Fitzpatrick, and Michael J Richardson. Measuring the dynamics of interactional synchrony. *Journal of Nonverbal Behavior*, 36(4):263–279, 2012.
- [41] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. v. d. Meij, and H. Hung. The matchn-mingle 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*, pages 1–1, 2018. ISSN 1949-3045. doi: 10.1109/TAFFC.2018.2848914.
- [42] Lu Zhang and Hayley Hung. Beyond f-formations: Determining social involvement in free standing conversing groups from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1086–1095, 2016.
- [43] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer, 2005.
- [44] Daniel Gatica-Perez, Jean-Marc Odobez, Sileye O Ba, Kevin C Smith, and Guillaume Lathoud. Tracking people in meetings with particles. Technical report, IDIAP, 2004.
- [45] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1707–1720, 2015.
- [46] Gloria Zen, Bruno Lepri, Elisa Ricci, and Oswald 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.
- [47] Alircza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. IEEE, 2012.
- [48] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015.

- [49] Cheng Chen, Alexandre Heili, and Jean-Marc Odobez. A joint estimation of head and body orientation cues in surveillance video. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 860–867. IEEE, 2011.
- [50] Elisa Ricci, Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Buló, Narendra Ahuja, and Oswald Lanz. Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4660–4668, 2015.
- [51] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10873–10883, 2019.
- [52] Sebastian Gorga and Kazuhiro Otsuka. Conversation scene analysis based on dynamic bayesian network and image-based gaze detection. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 1–8, 2010.
- [53] Sebastiano Vascon, Eyasu Z Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Computer Vision and Image Understanding*, 143:11–24, 2016.
- [54] Laura Cabrera-Quiros, David MJ Tax, and Hayley 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. *IEEE Transactions on Multimedia*, 22(1):138–147, 2019.
- [55] Cade McCall. Mapping social interactions: the science of proxemics. *Social behavior from rodents to humans*, pages 295–308, 2015.
- [56] Ron Dotsch and Daniël HJ Wigboldus. Virtual prejudice. *Journal of experimental social psychology*, 44(4):1194–1198, 2008.
- [57] Cade McCall, Jim Blascovich, Ariana Young, and Susan Persky. Proxemic behaviors as predictors of aggression towards black (but not white) males in an immersive virtual environment. *Social Influence*, 4(2):138–154, 2009.
- [58] Michael L Spezio, Ralph Adolphs, Robert SE Hurley, and Joseph Piven. Analysis of face gaze in autism using “bubbles”. *Neuropsychologia*, 45(1):144–151, 2007.
- [59] Sydney Thompson, Abhijit Gupta, Anjali W. Gupta, Austin Chen, and Marynel Vázquez. Conversational group detection with graph neural networks. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. ACM, October 2021. doi: 10.1145/3462244.3479963. URL <https://doi.org/10.1145/3462244.3479963>.

- [60] Mason Swofford, John Peruzzi, Nathan Tsoi, Sydney Thompson, Roberto Martín-Martín, Silvio Savarese, and Marynel Vázquez. 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.
- [61] Daniel Olguun Olguun and Alex Sandy Pentland. Sociometric badges: State of the art and future applications. In *Doctoral colloquium presented at IEEE 11th international symposium on wearable computers, Boston, MA*, 2007.
- [62] Matthew Carlson Dobson. Low-power epidemic communication in wireless ad hoc networks. 2013.
- [63] Oren Lederman, Dan Calacci, Angus MacMullen, Daniel C Fehder, Fiona E Murray, and Alex ‘Sandy’ Pentland. Open badges: A low-cost toolkit for measuring team communication and dynamics. *arXiv preprint arXiv:1710.01842*, 2017.
- [64] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5030–5039, 2018.
- [65] Jose Vargas Quiros, Stephanie Tan, Chirag Raman, Laura Cabrera-Quiros, and Hayley Hung. Covfee: an extensible web framework for continuous-time annotation of human behavior. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 265–293. PMLR, 2022.
- [66] Jeroen Lichtenauer, Jie Shen, Michel Valstar, and Maja Pantic. Cost-effective solution to synchronised audio-visual data capture using multiple sensors. *Image and Vision Computing*, 29(10):666–680, 2011.
- [67] Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, and Maja Pantic. The mahnob mimicry database: A database of naturalistic human interactions. *Pattern recognition letters*, 66:52–61, 2015.
- [68] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley 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.
- [69] Oren Lederman, Akshay Mohan, Dan Calacci, and Alex Sandy Pentland. Rhythm: A unified measurement platform for human organizations. *IEEE MultiMedia*, 25(1): 26–38, 2018.
- [70] Leonid Sigal, Alexandru O Balan, and Michael 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):4–27, 2010.

- [71] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.
- [72] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, pages 1–21. ISSN 0920-5691. URL <http://dx.doi.org/10.1007/s11263-012-0564-1>. 10.1007/s11263-012-0564-1.
- [73] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1):157–173, 2008.
- [74] Bruce W Tuckman and Mary Ann C Jensen. Stages of small-group development revisited. *Group & organization studies*, 2(4):419–427, 1977.
- [75] Rutger Rienks, Ronald Poppe, and Dirk Heylen. Differences in head orientation between speakers and listeners in multi-party conversations. *International Journal HCS*, 2005.
- [76] Dušan Jan and David R Traum. Dynamic movement and positioning of embodied agents in multiparty conversations. In *Proceedings of the Workshop on Embodied Language Processing*, pages 59–66. Association for Computational Linguistics, 2007.
- [77] Jari K Hietanen. Social attention orienting integrates visual information from head and body orientation. *Psychological Research*, 66(3):174–179, 2002.
- [78] Troy AW Visser and Ashton Roberts. Automaticity of social cues: The influence of limiting cognitive resources on head orientation cueing. *Scientific reports*, 8(1):10288, 2018.
- [79] Alex Pentland. *Social physics: How good ideas spread-the lessons from a new science*. Penguin, 2014.
- [80] Hayley Hung, Ekin Gedik, and Laura Cabrera Quiros. Complex conversational scene analysis using wearable sensors. In *Multimodal Behavior Analysis in the Wild*, pages 225–245. Elsevier, 2019.
- [81] Chirag Raman and Hayley 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.
- [82] Tanya L Chartrand and John A Bargh. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893, 1999.

- [83] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.
- [84] Carlos Segura, Cristian Canton-Ferrer, Alberto Abad, Josep R Casas, and Javier Hernando. Multimodal head orientation towards attention tracking in smartrooms. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 2, pages II–681. IEEE, 2007.
- [85] Alberto Abad, Carlos Segura, Climent Nadeu, and Javier Hernando. Audio-based approaches to head orientation estimation in a smart-room. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.
- [86] Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulò, Narendra Ahuja, Oswald Lanz, and Elisa Ricci. Joint estimation of human pose and conversational groups from social scenes. *International Journal of Computer Vision*, 126(2-4): 410–429, 2018.
- [87] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Gaowen Liu, Oswald Lanz, and Nicu Sebe. A multi-task learning framework for head pose estimation under target motion. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1070–1083, 2015.
- [88] Stephanie Tan, David MJ Tax, and Hayley Hung. Improving temporal interpolation of head and body pose using gaussian process regression in a matrix completion setting. In *Proceedings of the Group Interaction Frontiers in Technology*, pages 1–8. 2018.
- [89] Manon Kok, Jeroen D Hol, and Thomas B Schön. Using inertial sensors for position and orientation estimation. *arXiv preprint arXiv:1704.06053*, 2017.
- [90] Valérie Renaudin, Muhammad Haris Afzal, and Gérard Lachapelle. New method for magnetometers based orientation estimation. In *IEEE/ION Position, Location and Navigation Symposium*, pages 348–356. IEEE, 2010.
- [91] Daniel Vlasic, Rolf Adelsberger, Giovanni Vannucci, John Barnwell, Markus Gross, Wojciech Matusik, and Jovan Popović. Practical motion capture in everyday surroundings. *ACM transactions on graphics (TOG)*, 26(3):35–es, 2007.
- [92] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.
- [93] Timo Von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. Human pose estimation from video and imus. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1533–1547, 2016.
- [94] Gerard Pons-Moll, Andreas Baak, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bodo Rosenhahn. Multisensor-fusion for 3d full-body human motion capture. In

- 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 663–670. IEEE, 2010.
- [95] Chulhong Min, Akhil Mathur, Alessandro Montanari, and Fahim Kawsar. An early characterisation of wearing variability on motion signals for wearables. In *Proceedings of the 23rd International Symposium on Wearable Computers*, pages 166–168, 2019.
- [96] Andrea Ferlini, Alessandro Montanari, Cecilia Mascolo, and Robert Harle. Head motion tracking through in-ear wearables. 2019.
- [97] Taku Hachisu, Yadong Pan, Soichiro Matsuda, Baptiste Bourreau, and Kenji Suzuki. Facelooks: A smart headband for signaling face-to-face behavior. *Sensors*, 18(7):2066, 2018.
- [98] Alessandro Montanari, Zhao Tian, Elena Francu, Benjamin Lucas, Brian Jones, Xia Zhou, and Cecilia Mascolo. Measuring interaction proxemics with wearable light tags. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):25, 2018.
- [99] Norene Kelly. All the world’s a stage: what makes a wearable socially acceptable. *interactions*, 24(6):56–60, 2017.
- [100] Arran T Reader and Nicholas P Holmes. Examining ecological validity in social interaction: problems of visual fidelity, gaze, and social potential. *Culture and Brain*, 4(2):134–146, 2016.
- [101] Benjamin N Waber, Daniel Olguin Olguin, Taemie Kim, and Alex Pentland. Understanding organizational behavior with wearable sensing technology. *Available at SSRN 1263992*, 2008.
- [102] Yang Lu, Shujuan Yi, Nan Hou, Jingfu Zhu, and Tiemin Ma. Deep neural networks for head pose classification. In *Intelligent Control and Automation (WCICA), 2016 12th World Congress on*, pages 2787–2790. IEEE, 2016.
- [103] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1087–1096, 2019.
- [104] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2074–2083, 2018.
- [105] Chavdar Papazov, Tim K Marks, and Michael Jones. Real-time 3d head pose and facial landmark estimation from depth images using triangular surface patch features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4722–4730, 2015.

- [106] Thanarat Horprasert, Yaser Yacoob, and Larry S Davis. Computing 3-d head orientation from a monocular image sequence. In *Proceedings of the second international conference on automatic face and gesture recognition*, pages 242–247. IEEE, 1996.
- [107] Ying Wu and Kentaro Toyama. Wide-range, person-and illumination-insensitive head orientation estimation. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 183–188. IEEE, 2000.
- [108] Jing Xiao, Tsuyoshi Moriyama, Takeo Kanade, and Jeffrey F Cohn. Robust full-motion recovery of head by dynamic templates and re-registration techniques. *International Journal of Imaging Systems and Technology*, 13(1):85–94, 2003.
- [109] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.
- [110] Ben Benfold and Ian Reid. Colour invariant head pose classification in low resolution video. In *Proceedings of the 19th British Machine Vision Conference*, September 2008.
- [111] Irtiza Hasan, Francesco Setti, Theodore Tsesmelis, Alessio Del Bue, Fabio Galasso, and Marco Cristani. Mx-1stm: mixing tracklets and vislets to jointly forecast trajectories and head poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6067–6076, 2018.
- [112] Sergey Prokudin, Peter Gehler, and Sebastian Nowozin. Deep directional statistics: Pose estimation with uncertainty quantification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018.
- [113] Kazuhiro Otsuka, Keisuke Kasuga, and Martina Köhler. Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 191–199. ACM, 2018.
- [114] Lucas Beyer, Alexander Hermans, and Bastian Leibe. Biternion nets: Continuous head pose regression from discrete training labels. In *German Conference on Pattern Recognition*, pages 157–168. Springer, 2015.
- [115] Gerard E Grossman, R John Leigh, Larry A Abel, Douglas J Lanska, and SE Thurston. Frequency and velocity of rotational head perturbations during locomotion. *Experimental brain research*, 70(3):470–476, 1988.
- [116] James L Crowley, Patrick Reignier, and Sebastien Pesnel. Context aware vision using image-based active recognition. 2004.
- [117] Cheng Chen and Jean-Marc Odobez. We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1544–1551. IEEE, 2012.

- [118] Cheng Chen, Alexandre Heili, and Jean-Marc Odobez. Combined estimation of location and body pose in surveillance video. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, pages 5–10. IEEE, 2011.
- [119] I Kåsa. A circle fitting procedure and its error analysis. *IEEE Transactions on instrumentation and measurement*, (1):8–14, 1976.
- [120] Taemie Kim, Agnes Chang, and Alex Sandy Pentland. Enhancing organizational communication using sociometric badges. In *Proceedings of the 11th International Symposium on Wearable Computers (Submitted)*, 2007.
- [121] Ekin Gedik and Hayley Hung. Personalised models for speech detection from body movements using transductive parameter transfer. *Personal and Ubiquitous Computing*, 21(4):723–737, 2017.
- [122] Dominik Gusenbauer, Carsten Isert, and Jens Krösche. Self-contained indoor positioning on off-the-shelf mobile devices. In *2010 International Conference on Indoor Positioning and Indoor Navigation*, pages 1–9. IEEE, 2010.
- [123] Ramsey Faragher and Robert Harle. Location fingerprinting with bluetooth low energy beacons. *IEEE journal on Selected Areas in Communications*, 33(11):2418–2428, 2015.
- [124] Samer S Saab and Zahi S Nakad. A standalone rfid indoor positioning system using passive tags. *IEEE Transactions on Industrial Electronics*, 58(5):1961–1970, 2010.
- [125] Sileye O. Ba and Jean-Marc Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1):101–116, January 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2010.69. URL <https://doi.org/10.1109/TPAMI.2010.69>.
- [126] Sileye O. Ba, Hayley Hung, and Jean-Marc Odobez. Visual activity context for focus of attention estimation in dynamic meetings. In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, ICME’09*, page 1424–1427. IEEE Press, 2009. ISBN 9781424442904.
- [127] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*, 2018.
- [128] Chirag Raman, Jose Vargas-Quiros, Stephanie Tan, Ekin Gedik, Ashraful Islam, and Hayley Hung. Conflab: A rich multimodal multisensor dataset of free-standing social interactions in-the-wild, 2022. URL <https://arxiv.org/abs/2205.05177>.
- [129] Marco Cristani, Ramachandra Raghavendra, Alessio Del Bue, and Vittorio Murino. Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, 100:86–97, 2013.

- [130] Dan Bohus, Chit W Saw, and Eric Horvitz. Directions robot: in-the-wild experiences and lessons learned. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 637–644. Citeseer, 2014.
- [131] Marynel Vazquez, Elizabeth J. Carter, Braden McDorman, Jodi Forlizzi, Aaron Steinfeld, and Scott E. Hudson. 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*, pages 42–52. ACM Press. ISBN 978-1-4503-4336-7. doi: 10.1145/2909824.3020207. URL <http://dl.acm.org/citation.cfm?doid=2909824.3020207>.
- [132] Jorge Rios-Martinez, Anne Spalanzani, and Christian Laugier. From proxemics theory to socially-aware navigation: A survey. *International Journal of Social Robotics*, 7(2): 137–153, 2015.
- [133] Annica Kristoffersson, Silvia Coradeschi, Kerstin Severinson Eklundh, and Amy Loutfi. Towards measuring quality of interaction in mobile robotic telepresence using sociometric badges. *Paladyn, Journal of Behavioral Robotics*, 4(1):34–48, 2013.
- [134] Navin Raj Prabhu, Chirag Raman, and Hayley Hung. Defining and quantifying conversation quality in spontaneous interactions. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 196–205, 2020.
- [135] Edmund T Hall and Edward Twitchell Hall. *The hidden dimension*, volume 609. Anchor, 1966.
- [136] T Matthew Ciolek and Adam Kendon. Environment and the spatial arrangement of conversational encounters. *Sociological Inquiry*, 50(3-4):237–271, 1980.
- [137] Michael J Hove and Jane L Risen. It’s all in the timing: Interpersonal synchrony increases affiliation. *Social cognition*, 27(6):949–960, 2009.
- [138] Ilka H Gleibs, Neil Wilson, Geetha Reddy, and Caroline Catmur. Group dynamics in automatic imitation. *PloS one*, 11(9):e0162880, 2016.
- [139] Mariëlle Stel, Jim Blascovich, Cade McCall, Jessanne Mastop, Rick B Van Baaren, and Roos Vonk. Mimicking disliked others: Effects of a priori liking on the mimicry-liking link. *European Journal of Social Psychology*, 40(5):867–880, 2010.
- [140] Maria M Egbert. Schisming: The collaborative transformation from a single conversation to multiple conversations. *Research on Language and Social Interaction*, 30(1): 1–51, 1997.
- [141] Erving Goffman. *Interaction ritual: Essays on face-to-face interaction*. 1967.
- [142] Erving Goffman. On face-work: An analysis of ritual elements in social interaction. *Psychiatry*, 18(3):213–231, 1955.
- [143] Bruce W. Tuckman. Developmental sequence in small groups. *Psychological Bulletin*, 63(6):384–399, 1965. ISSN 0033-2909. doi: 10.1037/h0022100. URL <http://content.apa.org/journals/bul/63/6/384>.

- [144] Lucy Baehren. Saying “goodbye” to the conundrum of leave-taking: a cross-disciplinary review. *Humanities and Social Sciences Communications*, 9(1):1–13, 2022.
- [145] Hooman Hedayati, Daniel Szafir, and Sean Andrist. Recognizing f-formations in the open world. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, March 2019. doi: 10.1109/hri.2019.8673233. URL <https://doi.org/10.1109/hri.2019.8673233>.
- [146] Dan Bohus, Sean Andrist, and Eric Horvitz. A study in scene shaping: Adjusting f-formations in the wild. 2018.
- [147] Dan Bohus and Eric Horvitz. Dialog in the open world. In *Proceedings of the 2009 international conference on Multimodal interfaces - ICMI-MLMI '09*. ACM Press, 2009. doi: 10.1145/1647314.1647323. URL <https://doi.org/10.1145/1647314.1647323>.
- [148] Marynel Vázquez, Aaron Steinfeld, and Scott E Hudson. Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3010–3017. IEEE, 2015.
- [149] Joe Connolly, Nathan Tsoi, and Marynel Vázquez. Perceptions of conversational group membership based on robots' spatial positioning: Effects of embodiment. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, March 2021. doi: 10.1145/3434074.3447195. URL <https://doi.org/10.1145/3434074.3447195>.
- [150] Tian Gan, Yongkang Wong, Daqing Zhang, and Mohan S Kankanhalli. Temporal encoded f-formation system for social interaction detection. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 937–946, 2013.
- [151] Francesco Setti, Oswald Lanz, Roberta Ferrario, Vittorio Murino, and Marco Cristani. Multi-scale f-formation discovery for group detection. In *2013 IEEE International Conference on Image Processing*. IEEE, September 2013. doi: 10.1109/icip.2013.6738732. URL <https://doi.org/10.1109/icip.2013.6738732>.
- [152] Lu Zhang and Hayley Hung. On social involvement in mingling scenarios: Detecting associates of f-formations in still images. *IEEE Transactions on Affective Computing*, 12(1):165–176, 2018.
- [153] Viktor Schmuck and Oya Celiktutan. Growl: Group detection with link prediction. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.
- [154] Nicolai Marquardt, Robert Diaz-Marino, Sebastian Boring, and Saul Greenberg. The proximity toolkit. In *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*. ACM Press, 2011. doi: 10.1145/2047196.2047238. URL <https://doi.org/10.1145/2047196.2047238>.

- [155] Kleomenis Katevas, Katrin Hänsel, Richard Clegg, Ilias Leontiadis, Hamed Hadadi, and Laurissa Tokarchuk. Finding dory in the crowd. In *Proceedings of the 1st Workshop on Machine Learning on Edge in Sensor Systems - SenSys-ML 2019*. ACM Press, 2019. doi: 10.1145/3362743.3362959. URL <https://doi.org/10.1145/3362743.3362959>.
- [156] Alessandro Montanari, Zhao Tian, Elena Francu, Benjamin Lucas, Brian Jones, Xia Zhou, and Cecilia Mascolo. Measuring interaction proxemics with wearable light tags. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–30, March 2018. doi: 10.1145/3191757. URL <https://doi.org/10.1145/3191757>.
- [157] Oren Lederman, Akshay Mohan, Dan Calacci, and Alex Sandy Pentland. Rhythm: A unified measurement platform for human organizations. *IEEE MultiMedia*, 25(1):26–38, January 2018. doi: 10.1109/mmul.2018.112135958. URL <https://doi.org/10.1109/mmul.2018.112135958>.
- [158] Spcl midge badge, 2022. URL https://github.com/Jerzeek/spcl_midge_hardware.
- [159] Samuel Rota Bulò and Marcello Pelillo. Dominant-set clustering: A review. *European Journal of Operational Research*, 262(1):1–13, 2017.
- [160] Stephanie Tan, David M. J. Tax, and Hayley 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, March 2021. doi: 10.1145/3448122. URL <https://doi.org/10.1145/3448122>.
- [161] Öykü Kapcak, Jose Vargas-Quiros, and Hayley Hung. Estimating romantic, social, and sexual attraction by quantifying bodily coordination using wearable sensors. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 154–160. IEEE, 2019.
- [162] Jose David Vargas Quiros, Oyku Kapcak, Hayley Hung, and Laura 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.
- [163] Tetsuro Funato, Shinya Aoi, Nozomi Tomita, and Kazuo Tsuchiya. Smooth enlargement of human standing sway by instability due to weak reaction floor and noise. *Royal Society Open Science*, 3(1):150570, 2016.
- [164] Margaret Wilson and Thomas P Wilson. An oscillator model of the timing of turn-taking. *Psychonomic bulletin & review*, 12(6):957–968, 2005.
- [165] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

- [166] Hung-Hsuan Huang, Naoya Baba, and Yukiko Nakano. Making virtual conversational agent aware of the addressee of users' utterances in multi-user conversation using nonverbal information. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 401–408. ACM, 2011.
- [167] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. *arXiv*, 2018.
- [168] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [169] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [170] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1799–1807. Curran Associates, Inc., 2014.
- [171] Weiming Hu, Tieniu Tan, Liang Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(3):334–352, Aug 2004. ISSN 1094-6977. doi: 10.1109/TSMCC.2004.829274.
- [172] Nicolo Carissimi, Paolo Rota, Cigdem Beyan, and Vittorio Murino. Filling the gaps: Predicting missing joints of human poses using denoising autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.
- [173] Stephanie Tan, David MJ Tax, and Hayley 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, 2021.
- [174] Cristian Canton-Ferrer, Carlos Segura, Josep R Casas, Montse Pardas, and Javier Hernandez. Audiovisual head orientation estimation with particle filtering in multisensor scenarios. *EURASIP Journal on Advances in Signal Processing*, 2008:32, 2008.
- [175] Manon Kok and Thomas B Schön. A fast and robust algorithm for orientation estimation using inertial sensors. *IEEE Signal Processing Letters*, 26(11):1673–1677, 2019.
- [176] Benoît Massé, Silève Ba, and Radu Horaud. Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2711–2724, 2017.

- [177] M. Higger, M. Akcakaya, and D. Erdogmus. A robust fusion algorithm for sensor failure. *IEEE Signal Processing Letters*, 20(8):755–758, Aug 2013. ISSN 1070-9908. doi: 10.1109/LSP.2013.2266254.
- [178] Benjamin A Newman, Reuben M Aronson, Siddartha S Srinivasa, Kris Kitani, and Henny Admoni. HARMONIC: A multimodal dataset of assistive human-robot collaboration. *arXiv preprint arXiv:1807.11154*, 2018.
- [179] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.
- [180] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.
- [181] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. PifPaf: Composite fields for human pose estimation. *arXiv preprint arXiv:1903.06593*, 2019.
- [182] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020.
- [183] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339, 2013.
- [184] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020.
- [185] István Sárádi, Timm Linder, Kai Oliver Arras, and Bastian Leibe. Metrabs: Metric-scale truncation-robust heatmaps for absolute 3d human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):16–30, 2020.
- [186] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, pages 20–40. Springer, 2020.
- [187] Amir Ghodrati, Marco Pedersoli, and Tinne Tuytelaars. Is 2d information enough for viewpoint estimation? *Proceedings BMVC 2014*, pages 1–12, 2014.
- [188] Marton Veges and A Lőrincz. Temporal smoothing for 3d human pose estimation and localization for occluded people. In *International Conference on Neural Information Processing*, pages 557–568. Springer, 2020.
- [189] Leonid Sigal and Michael J Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2041–2048. IEEE, 2006.

- [190] Donghoon Lee, Ming-Hsuan Yang, and Songhwai Oh. Head and body orientation estimation using convolutional random projection forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [191] Diego Tosato, Mauro Spera, Marco Cristani, and Vittorio Murino. Characterizing humans on riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1972–1984, 2012.
- [192] Jinyoung Choi, Beom-Jin Lee, and Byoung-Tak Zhang. Human body orientation estimation using convolutional neural network. *arXiv preprint arXiv:1609.01984*, 2016.
- [193] Mudassar Raza, Zonghai Chen, Saeed-Ur Rehman, Peng Wang, and Peng Bao. Appearance based pedestrians’ head pose and body orientation estimation using deep learning. *Neurocomputing*, 272:647–659, 2018.
- [194] Chenyan Wu, Yukun Chen, Jiajia Luo, Che-Chun Su, Anuja Dawane, Bikramjot Hanzra, Zhuo Deng, Bilan Liu, James Z Wang, and Cheng-hao Kuo. Mebow: Monocular estimation of body orientation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3451–3461, 2020.
- [195] Gabriele Fanelli, Thibaut Weise, Juergen Gall, and Luc Van Gool. Real time head pose estimation from consumer depth cameras. In *Joint Pattern Recognition Symposium*, pages 101–110. Springer, 2011.
- [196] Fumito Shinmura, Daisuke Deguchi, Ichiro Ide, Hiroshi Murase, and Hironobu Fujiyoshi. Estimation of human orientation using coaxial rgb-depth images. In *VISAPP (2)*, pages 113–120, 2015.
- [197] Kaoruko Okuno, Takayoshi Yamashita, Hiroshi Fukui, Shuzo Noridomi, Koji Arata, Yuji Yamauchi, and Hironobu Fujiyoshi. Body posture and face orientation estimation by convolutional network with heterogeneous learning. In *2018 International Workshop on Advanced Image Technology (IWAIT)*, pages 1–4. IEEE, 2018.
- [198] Wu Liu, Yongdong Zhang, Sheng Tang, Jinhui Tang, Richang Hong, and Jintao Li. Accurate estimation of human body orientation from rgb-d sensors. *IEEE Transactions on cybernetics*, 43(5):1442–1452, 2013.
- [199] Daniel Laidig, Marco Caruso, Andrea Cereatti, and Thomas Seel. Broad—a benchmark for robust inertial orientation estimation. *Data*, 6(7):72, 2021.
- [200] Liang Hu, Yujie Tang, Zhipeng Zhou, and Wei Pan. Reinforcement learning for orientation estimation using inertial sensors with performance guarantee. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10243–10249. IEEE, 2021.
- [201] Hamad Ahmed and Muhammad Tahir. Improving the accuracy of human body orientation estimation with wearable IMU sensors. *IEEE Transactions on instrumentation and measurement*, 66(3):535–542, 2017.

- [202] Mitchell Webber and Raul Fernandez Rojas. Human activity recognition with accelerometer and gyroscope: A data fusion approach. *IEEE Sensors Journal*, 21(15): 16979–16989, 2021.
- [203] S. Bomma and N. M. Robertson. Joint classification of actions with matrix completion. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2766–2770, Sept 2015. doi: 10.1109/ICIP.2015.7351306.
- [204] Ricardo S. Cabral, Fernando Torre, Joao P. Costeira, and Alexandre Bernardino. Matrix completion for multi-label image classification. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 190–198. Curran Associates, Inc., 2011. URL <http://papers.nips.cc/paper/4419-matrix-completion-for-multi-label-image-classification.pdf>.
- [205] Andrew Goldberg, Ben Recht, Junming Xu, Robert Nowak, and Xiaojin Zhu. Transduction with matrix completion: Three birds with one stone. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 757–765. Curran Associates, Inc., 2010. URL <http://papers.nips.cc/paper/3932-transduction-with-matrix-completion-three-birds-with-one.pdf>.
- [206] E. J. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, May 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2044061.
- [207] F. Bachoc, F. Gamboa, J. M. Loubes, and N. Venet. A gaussian process regression model for distribution inputs. *IEEE Transactions on Information Theory*, pages 1–1, 2017. ISSN 0018-9448. doi: 10.1109/TIT.2017.2762322.
- [208] Christopher KI Williams. Prediction with gaussian processes: From linear regression to linear prediction and beyond. In *Learning in graphical models*, pages 599–621. Springer, 1998.
- [209] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.
- [210] Christopher J Paciorek and Mark J Schervish. Nonstationary covariance functions for Gaussian process regression. In *Advances in neural information processing systems*, pages 273–280, 2004.
- [211] Alexander Vandenberg-Rodes and Babak Shahbaba. Dependent matern processes for multivariate time series. *arXiv preprint arXiv:1502.03466*, 2015.
- [212] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

- [213] FG Mertens, A Ghosh, and LVE Koopmans. Statistical 21-cm signal separation via Gaussian process regression analysis. *Monthly Notices of the Royal Astronomical Society*, 478(3):3640–3652, 2018.
- [214] Elisa Ricci, Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulò, Narendra Ahuja, and Oswald Lanz. 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.
- [215] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.
- [216] Mattias Heldner and Jens Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.
- [217] Rivka Levitan, Agustín Gravano, and Julia 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.
- [218] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012.
- [219] Xun Cao, Yebin Liu, and Qionghai Dai. A flexible client-driven 3d tv system for real-time acquisition, transmission, and display of dynamic scenes. *EURASIP Journal on Advances in Signal Processing*, 2009:1–15, 2008.
- [220] Tomáš Svoboda, Hanspeter Hug, and Luc Van Gool. Viroom—low cost synchronized multicamera system and its self-calibration. In *Joint Pattern Recognition Symposium*, pages 515–522. Springer, 2002.
- [221] Tanzeem Choudhury and Alex Pentland. Sensing and modeling human networks using the sociometer. In *Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings.*, pages 216–222. IEEE, 2003.
- [222] Daniel Mark Wyatt et al. *Measuring and modeling networks of human social behavior*. University of Washington, 2010.
- [223] Daniel Olguín Olguín, Benjamin N Waber, Taemie Kim, Akshay Mohan, Koji Ara, and Alex Pentland. 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.
- [224] St 12-1:2014 - smpte standard - time and control code. *ST 12-1:2014*, pages 1–41, 2014.

- [225] David Mills, Jim Martin, Jack Burbank, and William Kasch. Network time protocol version 4: Protocol and algorithms specification. 2010.
- [226] David L Mills. Internet time synchronization: the network time protocol. *IEEE Transactions on communications*, 39(10):1482–1493, 1991.
- [227] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multi-modal 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.
- [228] David Bannach, Oliver Amft, and Paul 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.
- [229] Alessio Rosatelli, Ekin Gedik, and Hayley 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.
- [230] Aleksandar Matic, Venet Osmani, and Oscar Mayora-Ibarra. Analysis of social interactions through mobile phones. *Mobile Networks and Applications*, 17(6):808–819, 2012.
- [231] Mathew Laibowitz, Jonathan Gips, R AyIward, Alex Pentland, and Joseph A Paradiso. A sensor network for social dynamics. In *2006 5th International Conference on Information Processing in Sensor Networks*, pages 483–491. IEEE, 2006.
- [232] Jonathan Gips and Alex Pentland. Mapping human networks. In *Fourth Annual IEEE International Conference on Pervasive Computing and Communications (PERCOM’06)*, pages 10–pp. IEEE, 2006.
- [233] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one*, 5(7), 2010.
- [234] 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.
- [235] 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.
- [236] Evita-Stavroula Fotinea, Eleni Efthimiou, Athanasia-Lida Dimou, Theo Goulas, Panayotis Karioris, Angelika Peer, Petros Maragos, Costas Tzafestas, Iasonas Kokkinos, Klaus Hauer, 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.

- [237] Leonid Sigal, Alexandru O Balan, and Michael 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.
- [238] John Eidson and Kang 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.
- [239] David Mills. Clock discipline algorithm, 2014. URL <https://www.eecis.udel.edu/~mills/ntp/html/discipline.html>.
- [240] Go pro hero 7 black. <https://gopro.com/en/nl/shop/cameras/hero7-black/CHDHX-701-master.html>.
- [241] Plura ethernet to ltc convertor. <https://plurainc.com/wp-content/uploads/2019/03/eELCmanual.pdf>.
- [242] Timecode systems mini-basestation. URL <https://www.timecodesystems.com/wp-content/uploads/2016/08/Pulse-manual-Web-1.1-1.pdf>.
- [243] Timecode systems synbacpro. <https://www.timecodesystems.com/synbac-pro/>.
- [244] Michael Hopfengaertner. An open-source sensor platform for analysis of group dynamics. *arXiv preprint arXiv:1901.04977*, 2018.
- [245] R Duncan Luce et al. *Response times: Their role in inferring elementary mental organization*. Number 8. Oxford University Press on Demand, 1986.
- [246] Carol A Fowler, Julie M Brown, Laura Sabadini, and Jeffrey 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.
- [247] Marianne Sonnby-Borgström, Peter Jönsson, and Owe Svensson. Emotional empathy as related to mimicry reactions at different levels of information processing. *Journal of Nonverbal behavior*, 27(1):3–23, 2003.
- [248] Ken W Grant, Virginie van Wassenhove, and David Poeppel. Discrimination of auditory-visual synchrony. In *AVSP 2003-International Conference on Audio-Visual Speech Processing*, 2003.
- [249] Focusrite scarlett-2i2. URL <https://focusrite.com/en/usb-audio-interface/scarlett/scarlett-2i2>.
- [250] Bluetooth core specifications. <https://www.bluetooth.com/specifications/bluetooth-core-specification/>.
- [251] Ralf Steinmetz. Human perception of jitter and media synchronization. *IEEE Journal on selected Areas in Communications*, 14(1):61–72, 1996.

- [252] Edward Sazonov, Vidya Krishnamurthy, and Robert 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.
- [253] Peter Volgyesi, Abhishek Dubey, Timothy Krentz, Istvan Madari, Mary Metelko, and Gabor Karsai. Time synchronization services for low-cost fog computing applications. In *2017 International Symposium on Rapid System Prototyping (RSP)*, pages 57–63. IEEE, 2017.
- [254] Plura inc. rubidium series. URL <https://www.plurainc.com/solutions/timers/rubidium-series/>.
- [255] Bernd Dudzik, Simon Columbus, Tiffany Matej Hrkalic, Daniel Balliet, and Hayley 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.
- [256] William Fleeson. Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of personality*, 75:825–862, 8 2007. ISSN 0022-3506. doi: 10.1111/J.1467-6494.2007.00458.X. URL <https://pubmed.ncbi.nlm.nih.gov/17576360/>.
- [257] Jennifer G. La Guardia and Richard 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. ISSN 00223506. doi: 10.1111/j.1467-6494.2007.00473.x.
- [258] Judith A. Hall, Terrence G. Horgan, and Nora A. Murphy. Nonverbal communication. <https://doi.org/10.1146/annurev-psych-010418-103145>, 70:271–294, 1 2019. ISSN 15452085. doi: 10.1146/ANNUREV-PSYCH-010418-103145. URL <https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010418-103145>.
- [259] Katherine 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.
- [260] Chittaranjan Andrade. Internal, external, and ecological validity in research design, conduct, and evaluation. *Indian journal of psychological medicine*, 40(5):498–499, 2018.
- [261] Élise Labonte-LeMoyne, François Courtemanche, Marc Fredette, and Pierre-Majorique Léger. How wild is too wild: Lessons learned and recommendations for ecological validity in physiological computing research. In *PhyCS*, pages 123–130, 2018.
- [262] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. The matchn mingle 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.
- [263] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. Social interaction discovery by statistical analysis of f-formations. In Jesse Hoey, Stephen J. McKenna, and Emanuele Trucco, editors, *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>.
- [264] Madhumita Murgia. Who’s using your face? the ugly truth about facial recognition. *Financial Times*, 2019.
- [265] Chirag Raman, Stephanie Tan, and Hayley 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. ISBN 9781450379885. doi: 10.1145/3394171.3413697. URL <https://doi.org/10.1145/3394171.3413697>.
- [266] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xiangyu Zhang, Xinyu Zhou, Erjin Zhou, and Jian Sun. Learning delicate local representations for multi-person pose estimation. In *European Conference on Computer Vision*, pages 455–472. Springer, 2020.
- [267] Abeba Birhane and Vinay Uday 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>.
- [268] Sebastiano Vascon, Eyasu Zemene Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. A game-theoretic probabilistic approach for detecting conversational groups. In *Asian conference on computer vision*, pages 658–675. Springer, 2014.
- [269] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [270] Loris Bazzani, Marco Cristani, Diego Tosato, Michela Farenzena, Giulia Paggetti, Gloria Menegaz, and Vittorio Murino. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 30(2):115–127, 2013.
- [271] C. Cattuto, W. V. D. Broeck, A. Barrat, V. Colizza, J. Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLoS ONE*, 5, 2010.

- [272] Marion Hoffman, Per Block, Timon Elmer, and Christoph 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.
- [273] Martin Atzmueller and Florian 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.
- [274] Daniel Chaffin, Ralph Heidl, John R Hollenbeck, Michael Howe, Andrew Yu, Clay Voorhees, and Roger Calantone. The promise and perils of wearable sensors in organizational research. *Organizational Research Methods*, 20(1):3–31, 2017.
- [275] Alessio Rosatelli, Ekin Gedik, and Hayley 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.
- [276] Laura Cabrera-Quiros, David M.J. Tax, and Hayley 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.
- [277] J. V. Quiros and H. Hung. CNNs and Fisher Vectors for No-Audio Multimodal Speech Detection. In *MediaEval*, 2019.
- [278] University of york research data management. <https://www.york.ac.uk/library/info-for/researchers/data/sharing/access/>.
- [279] Utrecht university research data management. <https://www.uu.nl/en/research/research-data-management/guides/handling-personal-data>.
- [280] Laura Cabrera-Quiros and Hayley 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.
- [281] Laura Cabrera-Quiros and Hayley 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.
- [282] Hayley Hung, Chirag Raman, Ekin Gedik, Stephanie Tan, and Jose 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.
- [283] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently Scaling up Crowd-sourced Video Annotation. *International Journal of Computer Vision*, 101(1):184–204, 2013. ISSN 0920-5691. doi: 10.1007/s11263-012-0564-1.
- [284] Computer Vision Annotation Tool (CVAT).

- [285] Jose Vargas Quiros, Stephanie Tan, Chirag Raman, Laura Cabrera-Quiros, and Hayley Hung. Covfee: an extensible web framework for continuous-time annotation of human behavior. In Cristina Palmero, Julio C. S. Jacques Junior, Albert Clapés, Isabelle Guyon, Wei-Wei Tu, Thomas B. Moeslund, and Sergio Escalera, editors, *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>.
- [286] Daniel 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.
- [287] Hayley Hung and Daniel Gatica-Perez. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia*, 12(6):563–575, 2010.
- [288] Hayley Hung, Yan Huang, Gerald Friedland, and Daniel 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.
- [289] Cigdem Beyan, Muhammad Shahid, and Vittorio Murino. RealVAD: A Real-world Dataset and A Method for Voice Activity Detection by Body Motion Analysis. *x*, 9210(c):1–16, 2020. doi: 10.1109/tmm.2020.3007350.
- [290] Muhammad Shahid, Cigdem Beyan, and Vittorio 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.
- [291] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [292] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.
- [293] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. *arXiv:1405.0312 [cs]*, February 2015.
- [294] Jiaxing Shen, Oren Lederman, Jiannong Cao, Florian Berg, Shaojie Tang, and Alex Sandy Pentland. GINA: Group Gender Identification Using Privacy-Sensitive Audio Data. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2018-Novem:457–466, 2018. ISSN 15504786. doi: 10.1109/ICDM.2018.00061.
- [295] Pranay Gupta, Anirudh Thatipelli, Aditya Aggarwal, Shubh Maheshwari, Neel Trivedi, Sourav Das, and Ravi Kiran Sarvadevabhatla. Quo Vadis, Skeleton Action Recognition ? *arXiv:2007.02072 [cs]*, July 2020.

- [296] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline, 2016. URL <https://arxiv.org/abs/1611.06455>.
- [297] Zhiguang Wang, Weizhong Yan, and Tim 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.
- [298] Chang Wei Tan, Angus Dempster, Christoph Bergmeir, and Geoffrey I. Webb. Multi-rocket: Multiple pooling operators and transformations for fast and effective time series classification, 2021. URL <https://arxiv.org/abs/2102.00457>.
- [299] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, February 2019.
- [300] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. 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, October 2020. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2019.2916873.
- [301] Wen Dong, Bruno Lepri, Alessandro Cappelletti, Alex Sandy Pentland, Fabio Pianesi, and Massimo Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 271–278, 2007.
- [302] Julia Eberle, Karsten Stegmann, Alain Barrat, Frank Fischer, and Kristine 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.
- [303] Nigel 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.
- [304] Chirag Raman, Hayley Hung, and Marco Loog. Social Processes: Self-Supervised Meta-Learning over Conversational Groups for Forecasting Nonverbal Social Cues. *arXiv:2107.13576 [cs]*, July 2021.
- [305] Andrew Housholder, Jonathan Reaban, Aira Peregrino, Georgia Votta, and Tauheed Khan Mohd. Evaluating accuracy of the tobii eye tracker 5. In *International Conference on Intelligent Human Computer Interaction*, pages 379–390. Springer, 2021.
- [306] Diederick C Niehorster, Roy S Hessels, and Jeroen S Benjamins. Glassesviewer: Open-source software for viewing and analyzing data from the tobii pro glasses 2 eye tracker. *Behavior Research Methods*, 52(3):1244–1253, 2020.

- [307] Mengxi Jia, Xinhua Cheng, Yunpeng Zhai, Shijian Lu, Siwei Ma, Yonghong Tian, and Jian Zhang. Matching on sets: Conquer occluded person re-identification without alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1673–1681, 2021.
- [308] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. What i see is what you see: Joint attention learning for first and third person video co-analysis. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1358–1366, 2019.
- [309] Diako Mardanbegi, Tobias Langlotz, and Hans Gellersen. Resolving target ambiguity in 3d gaze interaction through vor depth estimation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.
- [310] Misbah Ahmad, Imran Ahmed, Kaleem Ullah, Ayesha Khattak, Awais Adnan, et al. Person detection from overhead view: a survey. *International Journal of Advanced Computer Science and Applications*, 10(4), 2019.
- [311] Mohd Ansari, Dushyant Kumar Singh, et al. Human detection techniques for real time surveillance: A comprehensive survey. *Multimedia Tools and Applications*, 80(6):8759–8808, 2021.
- [312] Markus Enzweiler and Dariu M Gavrilă. Monocular pedestrian detection: Survey and experiments. *IEEE transactions on pattern analysis and machine intelligence*, 31(12):2179–2195, 2008.
- [313] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. Visual tracking: An experimental survey. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1442–1468, 2013.
- [314] Yingkun Xu, Xiaolong Zhou, Shengyong Chen, and Fenfen Li. Deep learning for multiple object tracking: a survey. *IET Computer Vision*, 13(4):355–368, 2019.
- [315] Fabian Flohr, Madalin Dumitru-Guzu, Julian FP Kooij, and Dariu M Gavrilă. A probabilistic framework for joint pedestrian head and body orientation estimation. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):1872–1882, 2015.
- [316] Mikael Ovaska, Joni Kultanen, Teemu Autto, Joonas Uusnäkki, Antti Kariluoto, Joonas Himmanen, Mikko Virtaneva, Pasi Kaitila, and Pekka Abrahamsson. Deep neural network voice activity detector for downsampled audio data: An experiment report. *arXiv preprint arXiv:2108.05553*, 2021.
- [317] JA Kelso. The haken–kelso–bunz (hkb) model: from matter to movement to mind. *Biological Cybernetics*, 115(4):305–322, 2021.
- [318] Emmanuelle Tognoli, Mengsen Zhang, Armin Fuchs, Christopher Beetle, and JA Scott Kelso. Coordination dynamics: a foundation for understanding social behavior. *Frontiers in Human Neuroscience*, 14:317, 2020.
- [319] Li An. Modeling human decisions in coupled human and natural systems: Review of agent-based models. *Ecological modelling*, 229:25–36, 2012.

- [320] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.
- [321] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d’Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research: a report from the neurips 2019 reproducibility program. *Journal of Machine Learning Research*, 22, 2021.
- [322] German Barquero, Johnny Núñez, Sergio Escalera, Zhen Xu, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero. 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.
- [323] Chirag Raman, Hayley Hung, and Marco Loog. Why did this model forecast this future? closed-form temporal saliency towards causal explanations of probabilistic forecasts. *arXiv preprint arXiv:2206.00679*, 2022.
- [324] Hayley Hung, Gwenn Englebienne, and Jeroen 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.
- [325] OpenCV. Open source computer vision library. <https://github.com/opencv/opencv>, 2015.
- [326] Idiap multi camera calibration suite. <https://github.com/idiap/multicamera-calibration>.
- [327] Tdkicm20948. <https://invensense.tdk.com/products/motion-tracking/9-axis/icm-20948/>. Accessed: 2021-10-15.
- [328] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019.
- [329] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. 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.
- [330] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017.
- [331] Ignacio Oguiza. tsai - a state-of-the-art deep learning library for time series and sequential data. Github, 2022. URL <https://github.com/timeseriesAI/tsai>.

APPENDIX

Appendix for Chapter 6:

The Appendix is organized as follows:

- Hosting, licensing, and organization information for Conflab
- Documentation for Conflab, following Datasheets for Datasets [320]
- Sample post-hoc behavioral analysis report sent to each Conflab participant
- Details about out data-capture setup
- Implementation details for models used in our benchmark research tasks
- Additional experimental results and ablations
- Details for reproducibility following the ML Reproducibility Checklist [321]

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 1 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 [320] (see Appendix 7.5)
- 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

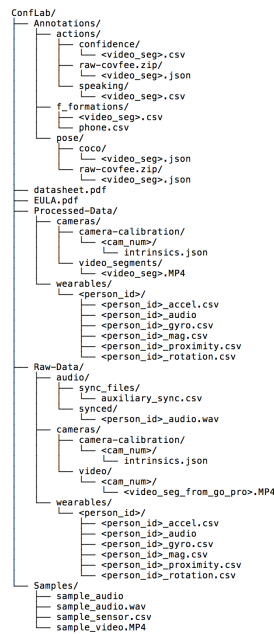


Figure 1: File structure of the Conflab dataset

- 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

B: DATASHEET FOR CONFLAB

This document is based on *Datasheets for Datasets* by Gebru *et al.* [320].

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 [267], 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 6.4 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 [265, Sec. 3.3]. This often requires modeling simplifications such as the summarizing of features over rolling windows [121, 276, 277]. 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 shows the structure of these instances and their relationships.

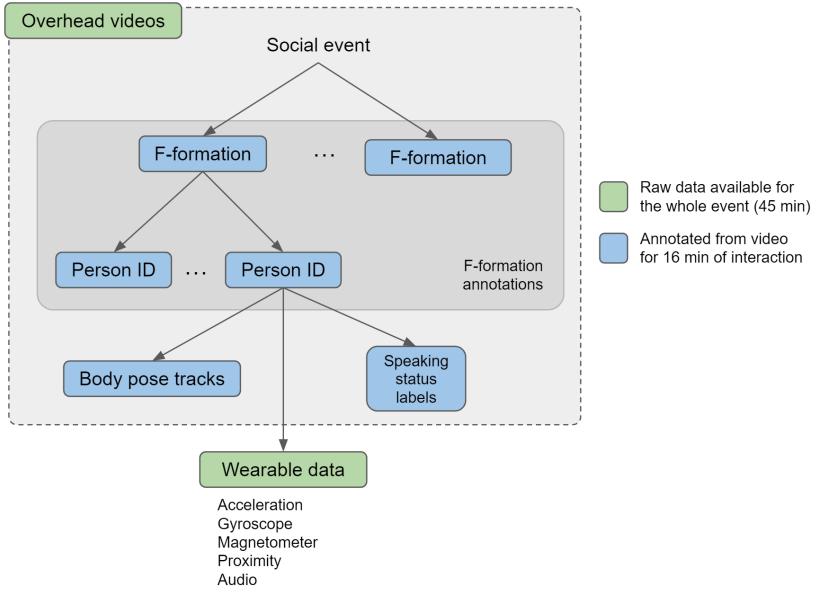


Figure 2: Structure of some of the instances in the dataset and their relationships. The interaction space was captured via overhead videos, in which f-formations (conversation groups) were annotated. An F-formation consists of set of people interacting for a variable period of time, and identified via 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.

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 × 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, 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.

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 [285].

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 [276].

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 6.5.

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. 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 7.5 of our associated paper) for information about the splits that we used in our 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?

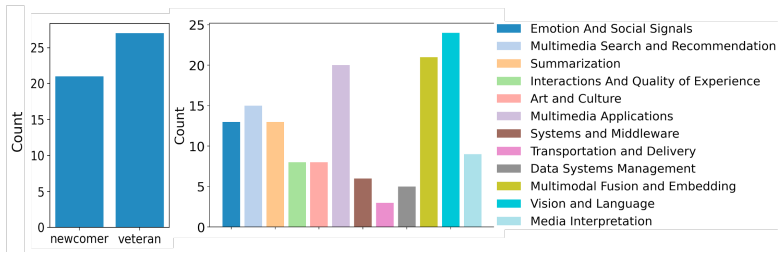


Figure 3: Distribution of participant seniority (left) and research interests (right) in percentage.

- 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 3 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.

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

Table 1: 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

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 multi-modal 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 [285], 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 [265], 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 4.

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

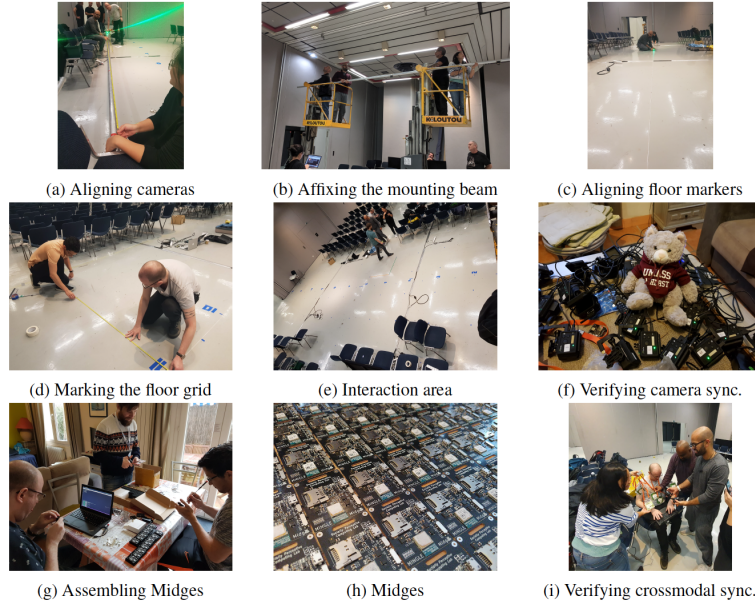


Figure 4: Illustrating the process of setting up the data recording.

The resources required to run this first edition of ConfLab include equipment, logistics, and travel costs. Table 1 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)?

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

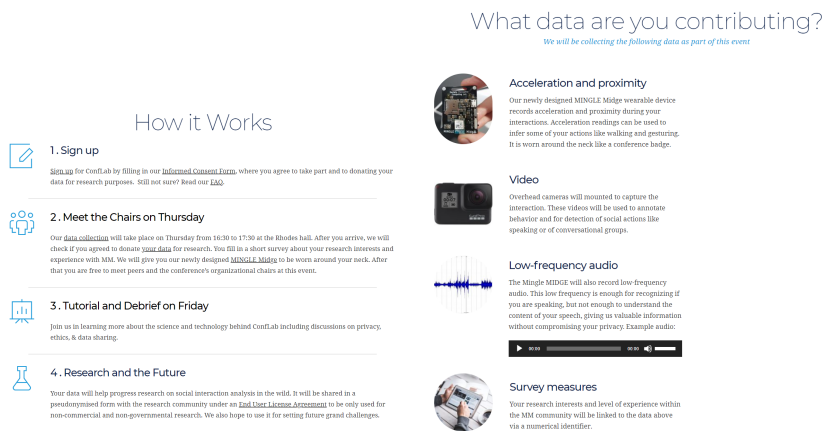


Figure 5: Screenshots of the Conflab web-page used for participant recruitment and registration.

the chairs and create a community dataset, the attendees also received a personalised report of their social behaviour from the wearable sensors (see Appendix 7.5) Conference 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)?

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 5 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 6.

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.

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.

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

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 [304, 322]. 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 [323].

Data Association. A crucial assumption made in many former multimodal datasets[45, 262, 269] is that the association of video data to the wearable modality can be manually performed. Few works [280, 281] 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 [81]. 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 [17]) 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 [262, 324].

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 [14, 134], dominance [288], rapport [35], and influence [301].

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 [162]) which need tolerances as low as 40 ms [265, Sec.3.2]. Prior works coped with lower tolerances by computing summary statistics over input windows [121, 276, 277]. 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 [299]. 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.

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.

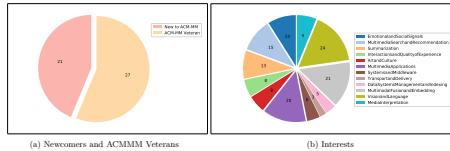


Figure 1: Statistics of Conflab participants

1

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.

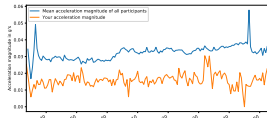


Figure 4: Acceleration magnitudes

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

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).

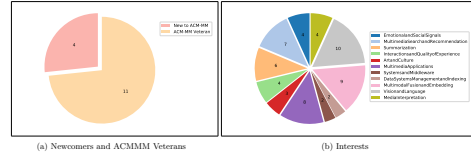


Figure 2: Statistics of people you interacted with

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.

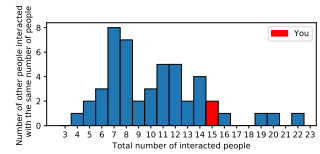


Figure 3: Distribution of the numbers of people participants interacted with

2

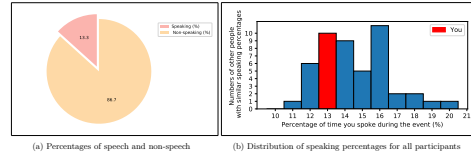


Figure 5: Your speaking behaviour

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

4

Figure 7: 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 6.3).

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 synchronization 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 [265, 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 $1\text{ m} \times 1\text{ m}$ 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 [325]. The calibration was performed using the Idiap multi camera calibration suite [326]. All wearable sensors include one TDK InvenSense ICM-20948 IMU [327] 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.

E: IMPLEMENTATION DETAILS

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, 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.

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 [125], 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.

F: ADDITIONAL RESULTS

PERSON AND KEYPOINTS DETECTION

Predictions from pretrained SOTA models Figure 8 shows predictions from SOTA human keypoint estimation models, namely, RSN [266], MSPN[328], HigherHRNet [329], and HourglassAENet [330], for the testing images of the Conflab dataset. Note that RSN

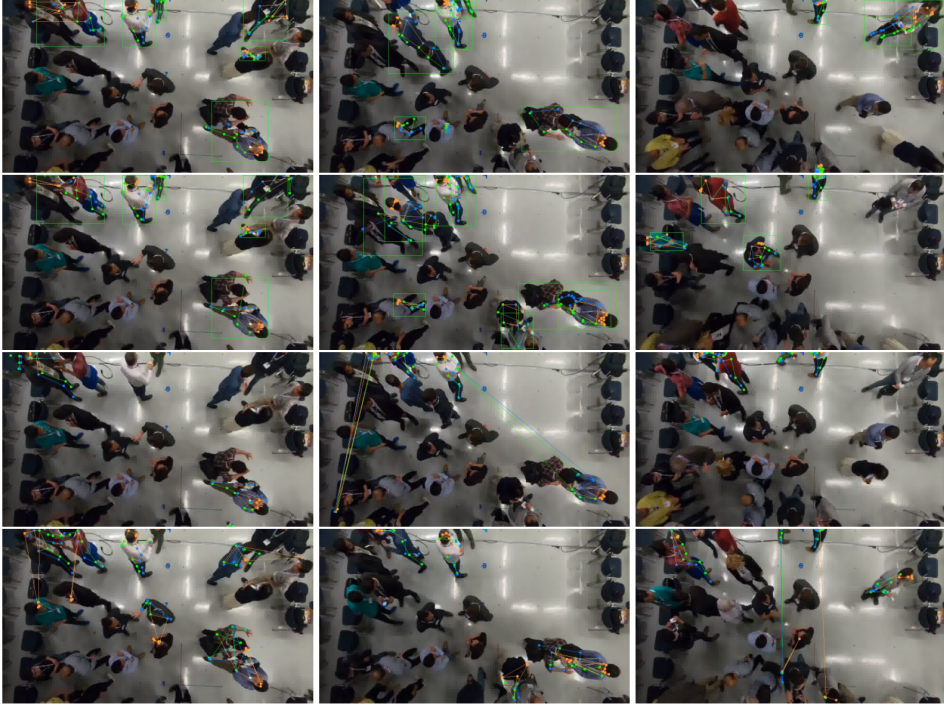


Figure 8: Results from Pretrained keypoint detection models. From top to bottom - predictions from RSN [266], MSPN[328], HigherHRNet [329], and HourglassAENet [330]. Results show that *SOTA 2D body keypoint detection models fail to capture the body keypoints in the Conflab dataset.*

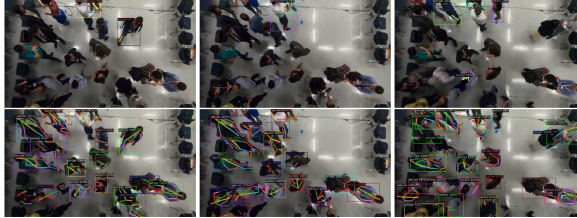


Figure 9: Results from (top) COCO pretrained Mask-RCNN model, (bottom) our Conflab finetuned Mask-RCNN model.

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.

Table 2: Effect of varying % frames from each camera at training on keypoint estimation.

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

Table 4: Keypoint estimation ablation with keypoints from different body sections: head and shoulders (5), + torso (9), + hips (11), + knees and feet (full 17).

#Keypoints	AP ₅₀ ^{OKS}	AP ^{OKS}	AP ₇₅ ^{OKS}
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 3: Effect of adding all frames from individual cameras to the training set on keypoint estimation.

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

Table 5: ROC AUC and accuracy for different sensor modalities from out 9-dof IMU in speaking status detection using the Minirocket classifier [298]. 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

Qualitative Results from ResNet-50 Finetuning Figure 9 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 and 3 include the results of our experiments investigating the effect of varying the training data size on keypoint detection performance (see main paper Section 6.6.1). In Table 4, 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 ($5 \rightarrow 9$, AP₅₀^{OKS} and AP^{OKS}), and that the relative gain when adding the hip keypoints ($9 \rightarrow 11$) is lower than when adding the lower body keypoints ($11 \rightarrow 17$, especially AP₇₅^{OKS}). We believe this is largely due to the lower body being more static relative to the arms that move a lot to execute gestures during conversations.

SPEAKING STATUS DETECTION

Experiments with different sensor modalities Table 5 displays the results from experiments using specific modalities from our IMUs for the task of speaking status detection. We used the best performing classifier (Minirocket [298]) among the ones tested in Table 6.2. The experiment setup is the same as detailed in Section 6.6.2, and the model is not changed between runs, except for the fact that different modalities may have a different number of input channels.

REPRODUCIBILITY CHECKLIST

PERSON AND KEYPOINTS DETECTION

- Source code link: <https://github.com/TUDeft-SPC-Lab/conflab>
- Data used for training: 112k frames (1809k person instances).
- Pre-processing: See Section 6.4, Appendix 7.5.
- 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.
- Hyperparameter 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 E
- Number of evaluation runs: 5
- How experiments were ran: See Section 6.6.1.
- Evaluation metrics: Average precision at different thresholds.
- Results: See Section 6.6.1 and Appendix E.
- Computing infrastructure used: All baseline experiments were ran on Nvidia V100 GPU (16GB) with IBM POWER9 Processor.

SPEAKING STATUS DETECTION

- Source code link: <https://github.com/TUDeft-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 6.6.2). 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 [331]. For the MS-G3D baseline [295], 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 6.6.2
- Computing infrastructure used: Experiments were ran on a personal computer with GPU acceleration (NVidia RTX3080).

F-FORMATION DETECTION

- Source code link: <https://github.com/TUdelft-SPC-Lab/conf1ab>
- Data used for training: Camera 6
- Pre-processing: See Section 7.5 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 6.3 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 6.6.3
- Computing infrastructure used: The experiments were run on Linux-based cluster instances on CPU with Matlab 2018a.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisors, Hayley and David, for your continuous support of this PhD thesis and research. Hayley, thank you for believing in me which is the only reason why I came to the Netherlands. These four years has been a bumpy but exciting ride, and I'm glad that we went through it together. Your enthusiasm for research and commitment to shape this community is truly inspiring. Thank you for your patience, motivation, and sharing your knowledge with me. David, thank you for your guidance and support throughout these years. By working with you, I truly realized the importance of questioning every result. You are a great example of upholding the highest level of scientific integrity. I would also like to thank my promoter, Marcel, for your support of this thesis. Your opinion is always on point and advice always practical. Members of the SPC group, Chirag, Jose, Bernd, Tiffany, Ekin, and Laura, it was a great pleasure knowing and working closely with you. I have learned a lot from our group meetings and reading groups. Being part of this interdisciplinary team has transformed me as a researcher, and reminded me to be curious and have an open mind. Chirag and Jose, thank you for being my MINGLE team mates for the majority of this journey. It is difficult to put down in words, but perhaps knowing that we have pulled through some of the hardest struggles together is always heart-warming for me. I'm extremely proud of what we achieved as a team. Thank you.

I thank the whole PRB group for creating a friendly, stimulating, and accommodating atmosphere for scientific discourse. The Monday meetings in Turing and coffee talks in Ritchie are among my favorite activities in my routine. Thank you for the conversations next to the coffee machines, in the alleys, and at the Borrels. It was a great pleasure and honor knowing all of you.

I would like to thank my family for their support. Sid, thank you for being my best friend and partner in life. You have been there since the beginning of this PhD journey which is now coming to an end. Being in academia yourself, you have always shared your personal and helpful viewpoint that guided me through the hardest times. Charleen, thank you for being the most supportive sister that one could ever ask for. Our daily calls of talking about the most stupid things kept me going, especially through the covid lockdown times. I'm extremely grateful to have you in my life. Jingnan, the quiet and super handy brother-in-law, you are my go-to person for anything about potatoes and start-ups. Thank you for taking so great care of the dogs. Dobby, thank you for being the funniest and goofiest dog that brings me the greatest joy. Miwa and Zaizai, thank you for being well-behaved siblings of Dobby. Lastly, I thank my parents, Jeffrey and Sharon, for relentless support since my Day 1. My upbringing has shaped me into who I am today, and will forever fuel my courage to face hardships and my passion for exploring the unknowns.

*Stephanie
Delft, November 2022*

CURRICULUM VITÆ

Stephanie TAN

1993/05/14	Date of birth in Pasadena, California, USA
2011-2015	BSc in chemical engineering California Institute of Technology (Caltech)
2016-2017	MSc in computing science Imperial College London
2018-2022	PhD in computer science Delft University of Technology (TU Delft)

LIST OF PUBLICATIONS

1. **Stephanie Tan**, David M.J. Tax, and Hayley Hung. **Conversation Group Detection With Spatio-Temporal Context**. 24th ACM International Conference on Multimodal Interaction (ICMI), 2022.
2. Chirag Raman*, Jose Vargas-Quiros*, **Stephanie Tan***, Ekin Gedik, Ashraful Islam, and Hayley Hung. "ConFLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild." To appear in 36th Conference on Neural Information Processing Systems (NeurIPS) Dataset and Benchmark Track, 2022.
3. **Stephanie Tan**, David M.J. Tax, and Hayley Hung. "Head and Body Orientation Estimation with Sparse Weak Labels in Free Standing Conversational Settings." Understanding Social Behavior in Dyadic and Small Group Interactions. Proceedings of Machine Learning Research (PMLR). 2022.
4. **Stephanie Tan**, David M.J. Tax, and Hayley 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 (IMWUT) 5.1 (2021): 1-22.
5. Chirag Raman*, **Stephanie Tan***, and Hayley Hung. "A modular approach for synchronized wireless multimodal multisensor data acquisition in highly dynamic social settings." Proceedings of the 28th ACM International Conference on Multimedia (ACM MM). 2020.
6. **Stephanie Tan**, David M.J. Tax, and Hayley Hung. "Improving temporal interpolation of head and body pose using Gaussian process regression in a matrix completion setting." Proceedings of the Group Interaction Frontiers in Technology. 2018. 1-8.

* equal contribution