Towards Automatic Social Involvement Estimation

Li, Zonghuan

DOI
10.1145/3678957.3688615

Publication date
2024

Document Version
Final published version

Published in
ICMI 2024 - Proceedings of the 26th International Conference on Multimodal Interaction

Towards Automatic Social Involvement Estimation

Zonghuan Li
Department of Intelligent Systems, EEMCS, TU Delft
Delft, The Netherlands
z.li-25@tudelft.nl

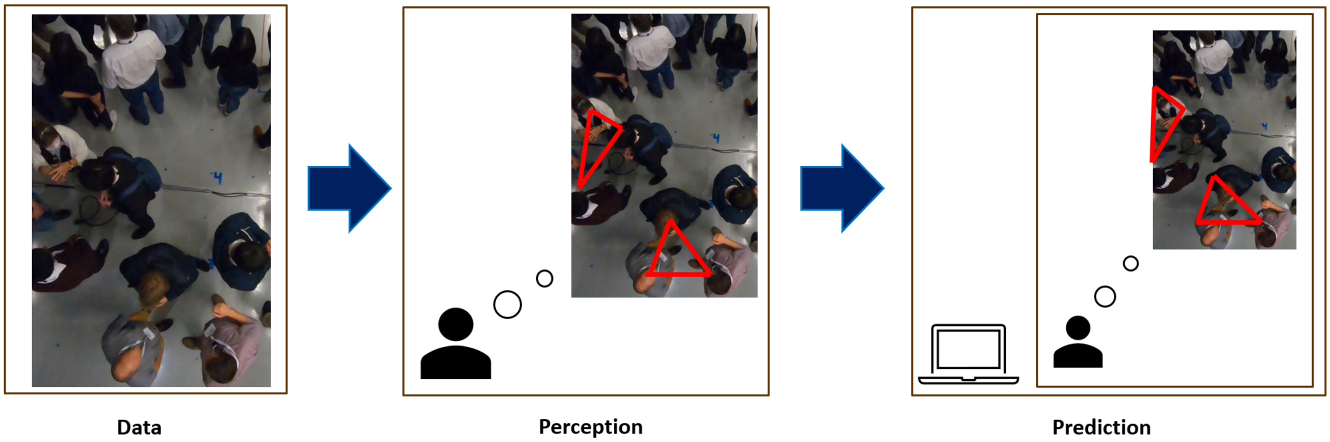

Figure 1: Model for research on social involvement estimation. Image taken from screenshot of the ConfLab [24] dataset. Human faces bulrred for privacy.

Abstract

Engaged in the group of Human Oriented Machine Intelligence(HOMI), my colleagues and I are working on the NEON project which focuses on social intention estimation. In particular, my job is to model and estimate social involvement, which is highly connected with the overall target. Studying social involvement will not only benefit our understanding of social interactions, but also provide crucial information for other potential applications. The work package of the research includes 1) build-up of a theoretical framework; 2) designing of the annotation framework; 3) collection of new dataset(s); 4) designing and training of the computer model.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Machine learning**.

Keywords

Social Involvement; Social Engagement; F-formation; Conversational Floors; Machine Learning

ACM Reference Format:
Zonghuan Li. 2024. Towards Automatic Social Involvement Estimation. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3678957.3688615

1 Introduction

Social well-being is valued highly in regard of personal health and social harmony. As a PhD candidate of Human Oriented Machine Intelligence(HOMI) group, I and my colleagues are working towards enhancing people's experience during social interactions. The main goal of the project is to model and estimate social intentions, since recognizing the intentions, both realized or unrealized, contributes significantly to acquiring social well-being.

To achieve this, multiple research directions has been proposed. The first is social intention estimation, which aims to model and capture people's intentions during conversations, and generate plausible narratives. The second is social involvement estimation, which focuses on modelling social involvements and participation. The third is conversational event detection, which mainly answers who is talking with whom given a conversational scene. Lastly, all members collaborate on collecting new datasets of such scenes.

It is worth noting that both social intention and social involvement are highly subjective, and people with different backgrounds are very likely to produce different narratives. This is why we focus on generating plausible descriptions of such scenarios instead of defining and finding an universal truth. This inherit subjectivity is also discussed and embodied in the following sections.

2 Research Direction

My research direction is about social involvement estimation in complex conversational scenes(CCS) [12]. Typically, I am trying

ICMI '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3688615

to develop an automated method to detect and predict involvements precisely in time and space. Given multi-modal data, for instance, videos, proximity, acceleration, etc, the aim is to label the involvement that are close to human perception.

2.1 Motivation

Modelling and estimating social involvement is of interest in two aspects, both descriptively and a practically. In a descriptive perspective, involvement can provide detailed description regarding social conversations. Since Kendon proposed the F-formation [14], it has become a wide accepted and utilized concept in social sciences and computer sciences. However, although F-formations illustrate the spatial organization of conversations, and Kendon did describe the dynamics of arrangement and maintenance of it, a more granular narrative of how each participant were attended and contributed to the conversation, is missing. Thus it becomes a natural inclination to study the intrinsic patterns of attention and participation displayed by each person during a conversation, and to develop both a theoretical framework and a computer model to describe it.

In a practical perspective, social involvement can provide important information for downstream tasks. For example, social intention estimation, as the joint goal of our group, can greatly benefit from the results of involvement estimation. It would help infer many subtle intentions that are not directly observable, such as intending to leave a conversation. Also, it could help evaluate the quality of an interaction, in the sense that one would expect the perceived quality of a participant to be high if they were highly involved in the interaction for most of the time, and vice versa. This also aligns with the overall target of out project to bring better social experience to people.

2.2 Research Question

The question of estimating social involvements in a conversational scene needs the answers of the following sub-questions:

Who is involved in the conversation? Of course, a preliminary question must be answered prior to this: *How many conversations are there in the scene and where are them?* This will be explored in the conversational event detection sub-project within the group. Suppose we have the answer, then this becomes determining participants that belong to the specified conversational group. In many occasions, conversational groups are represented as F-formations [14], but not necessarily, as some literature also mentions that there may be more than one conversational floor in an F-formation [23].

How involved is the person? As one could imagine, the levels of involvements are different among people, which is easily perceived by human when watching the videos. However, this could prove to difficult for computers, as at least two questions needs to be answered: 1)*What is the definition of involvement?* We need a theoretical framework that is both grounded in social sciences and operational in terms of implementing. 2) *What factors should be taken into consideration when estimation involvements?* In other words, we are considering the indicative and observable clues in the data. Beyond this, we also need a mathematical representation of involvement, which is discussed later.

3 Background and Related Work

3.1 F-formations

The problem of analyzing the structure of social conversations has been of interest in the intersection of multi-disciplinary studies. Adam Kendon has proposed the F-formation theory [14], where he defined the F-formation as

> F-formations arises when two or more people cooperate together to maintain a space between them to which they all have direct and exclusive access.

On top of this, Kendon also modelled the transactional segments of people, and described the *r-space, p-space, and o-space* of an F-formation.

In computer sciences, this definition has been widely adopted for many tasks recognizing F-formations. In these practices, F-formations are often mixed used with the term *conversational groups*. This, however, may not be precise as we disccus in section 3.3

Several methods has been proposed to detect F-formations in a rather straightforward way. [5] used voting in Hough space to determine the centers of F-formations. Based on graphs, [13] used maximum cliques to cluster conversational groups.

Some other methods consider this in a multi-modal approach. Yet, many methods chose not to use video or audio data, such as [11] and [9], where they proposed to detect conversation groups using an accelerometer. The former implemented the detection based on analyzing each person's actions, and the latter claimed that besides commonly used proximity data of individuals, the dynamics between people in interaction is also important and indicative. Also, in [25] and [30], the authors claimed that traditional methods of estimating head orientation has limitations and proposed a joint approach on proximity and interpersonal dynamics.

Other models, however, chose to leverage video or audio data. For instance, [27] and [36] detected F-formations in images and videos. The former modelled each person's transactional segment and optimized the posterior probability for assigning people to a group, while the latter modelled the frustums and predicted F-formations using a game-theoretical approach. Other methods utilized neural networks in their structures, such as [29, 31, 33], where MLP, GNN, and LSTM are used respectively. It is worth noting that a large part of these methods formulated an affinity matrix after processing the features, and applied the Dominant Set algorithm proposed by [13] to predict the F-formations, showing the effectiveness of this approach.

3.2 Social Involvement

Social involvement is central to my PhD research. It is worth noting here that in many literature this concept is not distinguished with the term of social engagement. In the following definitions some authors described social involvements, while others elaborated on social engagement. They are listed together and both regarded as social involvement in our context. The two terms are mixed used in this text.

Regarding this, many definitions of involvement has been proposed. In Tannen's book *Talking Voices* [32] the author summarized several viewpoints of involvements. 1) Gumperz [10] defined social involvement as the result of inferring the overall intention globally and the meaning of individual words locally. This also resonates with what Goodwin [8] gave. 2) On the other hand, Chafe [18] has claimed that involvement is more of a "psychological, internal state which shows itself in observable linguistic phenomena", which is closer to the author's understanding.

Sydner et al. [28] defined engagement as "*The process by which two (or more) participants establish, maintain and end their perceived connection*", in whichengagement was modelled as a process. Some other authors, like Poggi et al [21], modelled engagement as the value in which the participant contributes to maintain the interaction. This is referring engagement as a numerical value, and similar to what Peters et al [4] gives, modelling involvement as a status. On top of this, it is feasible to model the level of engagement of person p at time t as mathematical terms like $E_p(t)$. A more detailed review can be found in Oertel's work [16] as well as pellet's [19].

3.3 Conversational Floors and Schisming

As stated in [23], F-formations sometimes contain more than one conversational floor. [17] also pointed out that not all participants during an interaction are equally concentrated. Therefore, it is natural to consider the more granular structures of a conversation, especially in terms of the speaking. Conversational floors carry information about speaking turns. By Edelsky [6], conversational floors are defined as "the acknowledged what's-going-on within a psychological time/space". However, this is referring to the mental status of a person and is not a straightforwardly observable clue. Some other scholars, like McNeil [15], are using similar terms such as Growth Point to detect the dynamics of discourse.

[7] studied schisming, which is also of great interest. Schisming refers to the phenomenon of "conversation splits up into two or more conversations". This is important because 1) it happens frequently in almost every conversational scene; 2) it describes the point of change in conversational floors, and is therefore indicative of attention and involvements; 3) it relates conversational floors with spatial organizations, as the author studied the influence of eye gazes, body orientations and lexical contents.

4 Research Plan and Methodology

My research plan mainly consists of four parts: theoretical framework, annotation framework, modelling, and dataset collection. Each of the tasks are not independent from each other, and they are not divided in such an order to be completed one after the other. Rather, the work package should be done in a circular way, and each step shall be reflected by the results from other tasks.

4.1 Theoretical Framework

Currently, a unified framework is missing to describe the overall structure and dynamics of conversations, especially regarding capturing F-formations, conversational involvements, and conversational floors jointly. Importantly, it should state clearly the definition of social involvement and its connection with other concepts, such as intentions and gestures.

This framework needs to be developed by doing a thorough literature search on related topics, including F-formations and conversational floors. A well-known protocol, namely PRISMA [1], will be used for doing this review.

The framework will be used to provide operational criteria for designing the annotation framework. How we define and describe involvements will be embodied in the annotation guidelines, as well as the design of model architectures.

Another crucial effect of this framework is to provide the mathematical representation of involvements. Currently this can be discussed in several aspects. 1) Nature. As one might think initially, involvement could be defined based on each person [21, 28], but as pointed out in [37], it might be easier for annotators to label involvements by comparison. That being said, the annotations might be more reliable and intuitive if they represent "involvement between two or more people" instead of "involvement of one person". 2) Subject. In traditional settings, typically they assume that only one F-formation or conversation floor existed, and identifying the context of involvement was rather easy. However, in CCS where there are multiple of these interaction units, we first need to locate the conversation, namely, which F-formation or conversation floor is the focused conversation. 3) Quantitative description. For example, some people may think involvement as a level that can be described by a number between 0 and 1, but annotators may find it difficult to give a continuous evaluation. As described in [37], the description of emotional elements may be more of ordinal instead of interval or nominal. Furthermore, as pointed out in [26], there may be more than one dimension of involvement, as they considered both attention and emotion components. 4) Dimension. Some studies has proposed to use more than one component to describe involvement, such as [20] which included an interest and an emotion facet.

The difference in this representation will eventually affect the code implementation and objective functions in training. Therefore, this framework should be regarded as the basis for further steps, but it will also be reflected by the experimental results.

4.2 Dataset Collection

Currently there has been many datasets of conversational scenes, including ConfLab [24], MatchNMingle [3], Cocktail Party [38], Salsa [2], etc. However, the richness of modalities in datasets is never enough. We are planning to conduct a new dataset collection in a conference or similar setting, capturing videos, high and low frequency audios, proximity, acceleration, etc. I will work closely together with my colleagues to make sure that all collecting equipment are operational, and the whole process are compliant with GDPR. All participants will sign a consent form for collecting their information, and they will receive exact instructions during the collection process.

This step is expected to be done in the first or second year of my PhD. The collection should produce raw data for the annotators, although not all modalities needs to be annotated. Typically only video and audio will be assigned labels.

[1] https://www.prisma-statement.org/

4.3 Annotation Framework

An annotation framework is needed to give the annotators instructions and produce annotations for the data. In general, the recruited annotators will be given some basic knowledge about involvement, and provided instruction that help them state their perceptions best. This is crucial in our project, as the inner subjectivity of the labels must be taken into account. It is important that we design the guidelines in a way such that they can both understand the rules thoroughly, as well as not inhibiting expressing their first impressions.

Steps and criteria for choosing the annotators shall also be considered. If we use lay people, the advantage is that they are easier to recruit, but they may underperform by some complex instructions. Experts, or people with some knowledge in this area, may understand our tasks better, but they would be harder to find and more expensive on the compensations. Essentially, this is a practical matter, but would still impact on the quality of the training data.

The output of the annotation framework is the annotated datasets. It will be used for training the model. This step will also be reflected by the results of the model, as appropriately capturing the subjectivity is one crucial part in my research.

4.4 Baseline Investigation and Modelling

Before designing new models, it is worth considering how existing ones, especially recently popular Large Language Models(LLM) should play their roles in the estimation process. As one could imagine, many models such as GPT-4 [1] and Llama [35], are exhibiting very competitive abilities in analyzing social scenes, and one might ask why not simply adopt one of them. However, considering that not all LLMs are open weighted, people may have concern applying them to privacy sensitive data. Also, since LLMs are typically very large, they may not be deployed on small personal devices. Therefore, it is important to consider benefits and drawbacks of LLMs and decide what our models can contribute. A thorough investigation of potential applicable models, including LLMs, is needed to evaluate their performances and motivate novel designing of architectures.

To build a computer model to automatically analyze involvements in the datasets. Machine learning methods are expected to be used, and modern network structures will be utilized to optimize for performance. Among them, Graphical Neural Networks(GNN) is of special interest, as a graph denoted by vertices and edges $G = (V, E)$ is a natural and popular representation of groups and relationships. Novel architectures will be developed to adapt to the dataset, and will also be inspired by the theory of involvements. As done in typical computer machine learning, the training of the model is done by inputting training data and optimizing the target function, which is designed based on the selected metrics. After training, evaluation is performed to assess the performance of the model, and hyper-parameters are modified to seek the optimized configuration.

The evaluation of the model should be based on common metrics for machine learning, such as confusion matrices, AUC, etc. The model is expected to take in multi-model data, and output predicted or estimated involvements. We would also expect it to be ecologically valid and robust.

4.5 Metric Design

Metric is how we evaluate the model and determine the criteria of success. In general, we believe the criteria for the output (estimated involvement in CCS) should at least include: 1) Accuracy. This includes everything related with accuracy, such as precision, recall, AUC, etc. For classification, this refers to the correspondence with labels; for prediction, this refers to the correspondence with future observations. No subjectivity is included here; 2) Granularity. This refers to how fine-grained we can tell with the data. For example, can we estimate involvement in real-time, or can we only say something regarding to a time window/video segment? In most cases, the former is preferred; 3) Plausibility. This refers to how the results are aligned with the subjective narrative from the annotators. As there is no strict true or false, we may need to search for keyword/fact matching or emotion/attitude comparison. A process of designing is needed for this aspect.

5 Results and Remaining Work

Currently, I have identified some existing definitions of involvement, and investigated their applicable scenarios. I have also done preliminary experiments on the performances of LLMs, which showed the indication that they are rather unlikely to perform well in understanding involvement in CCS. In the next period, my focus would be on capturing eye gazes and leveraging the temporal dynamics of social interaction. On the one hand, eye gazes has been recognized as one of the most indicative factors of attention and involvements [22, 28]; on the other, many methods that detect F-formations, like [27] and [36], does not rely on the temporal continuity between frames. It is natural to think that involvements in conversational scenes, represented in videos and audios, are continuous in time and the temporal correlation could be used to improve the detection and estimation. Many modern video models, such as the [34], are capable of capturing this feature.

Also, I am now contributing to the firmware codes that are used to operate the sensors for data collection. Since we are planning an event with 30 to 50 people, it is non-trivial to monitor all sensors simultaneously and ensure that all of them work properly. Also, the alignment between modalities is crucial for the model training, which also needs to be guaranteed during collection [23].

6 Research Contributions

This research is expected to be a sub-project of the NEON. Understanding of social involvement will help us model and estimate social intention better, and eventually enhance social well-being. On its own, the theoretical framework may provide a novel and unified description of social involvements and benefit social science research. The computer model is an advanced way of automatically estimating social involvements, which not only assists understanding participation, attention during an interaction, but also serves potential applications in commercial, engineering, and social areas.

Acknowledgments

This paper is funded by ERC Consolidator Grant NEON: 1010891.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. 2015. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE transactions on pattern analysis and machine intelligence* 38, 8 (2015), 1707–1720.

[3] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. 2018. The MatchNMingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing* 12, 1 (2018), 113–130.

[4] Lee J Corrigan, Christina Basedow, Dennis Küster, Arvid Kappas, Christopher Peters, and Ginevra Castellano. 2015. Perception matters! Engagement in task orientated social robotics. In *2015 24th ieee international symposium on robot and human interactive communication (ro-man)*. IEEE, 375–380.

[5] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. 2011. Social interaction discovery by statistical analysis of f-formations.. In *BMVC*, Vol. 2. 10–5244.

[6] Carole Edelsky. 1981. Who's got the floor? *Language in society* 10, 3 (1981), 383–421.

[7] Maria M Egbert. 1997. Schisming: The collaborative transformation from a single conversation to multiple conversations. *Research on Language and Social Interaction* 30, 1 (1997), 1–51.

[8] Frederick D Erickson. 1984. Conversational Organization: Interaction between Speakers and Hearers.

[9] Ekin Gedik and Hayley Hung. 2018. 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 (2018), 1–24.

[10] John J Gumperz. 1982. *Discourse strategies*. Number 1. Cambridge University Press.

[11] Hayley Hung, Gwenn Englebienne, and Laura Cabrera Quiros. 2014. Detecting conversing groups with a single worn accelerometer. In *Proceedings of the 16th international conference on multimodal interaction*. 84–91.

[12] Hayley Hung, Ekin Gedik, and Laura Cabrera Quiros. 2019. Chapter 11 - Complex conversational scene analysis using wearable sensors. In *Multimodal Behavior Analysis in the Wild*, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe (Eds.). Academic Press, 225–245. https://doi.org/10.1016/B978-0-12-814601-9.00019-5

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

[14] Adam Kendon. 1990. *Conducting interaction: Patterns of behavior in focused encounters*. Vol. 7. CUP Archive.

[15] David McNeill, Susan Duncan, Amy Franklin, James Goss, Irene Kimbara, Fey Parrill, Haleema Welji, Lei Chen, Mary Harper, Francis Quek, et al. 2009. Mind merging. In *Expressing Oneself/Expressing One's Self*. Psychology Press, 143–164.

[16] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI* 7 (2020), 92.

[17] Catharine Oertel, Kenneth A Funes Mora, Joakim Gustafson, and Jean-Marc Odobez. 2015. Deciphering the silent participant: On the use of audio-visual cues for the classification of listener categories in group discussions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 107–114.

[18] David R Olson, Nancy Torrance, and Angela Hildyard. 1985. *Literacy, language and learning: The nature and consequences of reading and writing*. CUP Archive.

[19] Arthur Pellet-Rostaing, Roxane Bertrand, Auriane Boudin, Stéphane Rauzy, and Philippe Blache. 2023. A multimodal approach for modeling engagement in conversation. *Frontiers in Computer Science* 5 (2023), 1062342.

[20] Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi. 2005. A model of attention and interest using gaze behavior. In *International Workshop on Intelligent Virtual Agents*. Springer, 229–240.

[21] Isabella Poggi. 2006. Social influence through face, hands, and body. In *Second Nordic Conference on Multimodality Goteborg, Sweden*. Citeseer, 5–29.

[22] Isabella Poggi. 2013. *40. Mind, hands, face, and body: A sketch of a goal and belief view of multimodal communication.* De Gruyter Mouton, Berlin, Boston, 627–647. https://doi.org/doi:10.1515/9783110261318.627

[23] CA Raman. 2023. Towards Artificial Social Intelligence in the Wild: Sensing, Synthesizing, Modeling, and Perceiving Nonverbal Social Human Behavior. (2023).

[24] Chirag Raman, Jose Vargas Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and Hayley Hung. 2022. ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild. *Advances in Neural Information Processing Systems* 35 (2022), 23701–23715.

[25] Alessio Rosatelli, Ekin Gedik, and Hayley Hung. 2019. 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)*. IEEE, 147–153.

[26] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th international conference on Human-robot interaction*. 305–312.

[27] Francesco Setti, Chris Russell, Chiara Bassetti, and Marco Cristani. 2015. F-formation detection: Individuating free-standing conversational groups in images. *PloS one* 10, 5 (2015), e0123783.

[28] Candace L Sidner, Christopher Lee, and Neal Lesh. 2003. Engagement when looking: behaviors for robots when collaborating with people. In *Diabruck: Proceedings of the 7th workshop on the Semantic and Pragmatics of Dialogue*. Citeseer, 123–130.

[29] Mason Swofford, John Peruzzi, Nathan Tsoi, Sydney Thompson, Roberto Martín-Martín, Silvio Savarese, and Marynel Vázquez. 2020. Improving social awareness through dante: Deep affinity network for clustering conversational interactants. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[30] Stephanie Tan, David MJ Tax, and Hayley Hung. 2021. 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.

[31] Stephanie Tan, David MJ Tax, and Hayley Hung. 2022. Conversation group detection with spatio-temporal context. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 170–180.

[32] Deborah Tannen. 2007. *Talking voices: Repetition, dialogue, and imagery in conversational discourse*. Vol. 26. Cambridge University Press.

[33] Sydney Thompson, Abhijit Gupta, Anjali W Gupta, Austin Chen, and Marynel Vázquez. 2021. Conversational group detection with graph neural networks. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 248–252.

[34] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* 35 (2022), 10078–10093.

[35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[36] Sebastiano Vascon, Eyasu Z Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. 2016. Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Computer Vision and Image Understanding* 143 (2016), 11–24.

[37] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. 2017. The ordinal nature of emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 248–255.

[38] Gloria Zen, Bruno Lepri, Elisa Ricci, and Oswald Lanz. 2010. Space speaks: towards socially and personality aware visual surveillance. In *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*. 37–42.