Delft University of Technology

Multimodal Joint Head Orientation Estimation in Interacting Groups via Proxemics and Interaction Dynamics

Tan, Stephanie; Tax, David M.J.; Hung, Hayley

DOI
10.1145/3448122

Publication date
2021

Document Version
Final published version

Published in
Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies

Citation (APA)
Tan, S., Tax, D. M. J., & Hung, H. (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), Article 35. https://doi.org/10.1145/3448122

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.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Multimodal Joint Head Orientation Estimation in Interacting Groups via Proxemics and Interaction Dynamics

STEPHANIE TAN, DAVID M.J. TAX, and HAYLEY HUNG, Delft University of Technology, Netherlands

Human head orientation estimation has been of interest because head orientation serves as a cue to directed social attention. Most existing approaches rely on visual and high-fidelity sensor inputs and deep learning strategies that do not consider the social context of unstructured and crowded mingling scenarios. We show that alternative inputs, like speaking status, body location, orientation, and acceleration contribute towards head orientation estimation. These are especially useful in crowded and in-the-wild settings where visual features are either uninformative due to occlusions or prohibitive to acquire due to physical space limitations and concerns of ecological validity. We argue that head orientation estimation in such social settings needs to account for the physically evolving interaction space formed by all the individuals in the group. To this end, we propose an LSTM-based head orientation estimation method that combines the hidden representations of the group members. Our framework jointly predicts head orientations of all group members and is applicable to groups of different sizes. We explain the contribution of different modalities to model performance in head orientation estimation. The proposed model outperforms baseline methods that do not explicitly consider the group context, and generalizes to an unseen dataset from a different social event.

CCS Concepts: • **Computing methodologies → Neural networks**; **Supervised learning by regression**; **Temporal reasoning**;

Additional Key Words and Phrases: head orientation estimation, interaction dynamics, scene understanding

ACM Reference Format:
Stephanie Tan, David M.J. Tax, and Hayley Hung. 2021. Multimodal Joint Head Orientation Estimation in Interacting Groups via Proxemics and Interaction Dynamics. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 35 (March 2021), 22 pages. https://doi.org/10.1145/3448122

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 [53], when a new participant joins the conversation [28], or when there is a change of head and body orientation of other participants [25]. Head orientations, as a proxy for direction of attention [35, 61], can be reflective of the participant's mental processes and therefore, the interaction quality which is valuable information towards social intelligence [46]. 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.,

Authors' address: Stephanie Tan, s.tan-1@tudelft.nl; David M.J. Tax, D.M.J.Tax@tudelft.nl; Hayley Hung, h.hung@tudelft.nl, Delft University of Technology, Delft, Netherlands.

This work is licensed under a Creative Commons Attribution International 4.0 License.
© 2021 Copyright held by the owner/author(s).
2474-9567/2021/3-ART35
https://doi.org/10.1145/3448122

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 [5, 39, 44], e.g., seated meetings (Figure 1(a)), have been studied extensively, a closer analysis of complex conversational scenes [27], e.g., networking events or cocktail parties [57] (Figure 1(b) and 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 [57]. 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 [49]. 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 [20, 32]. 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 [11]. 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.

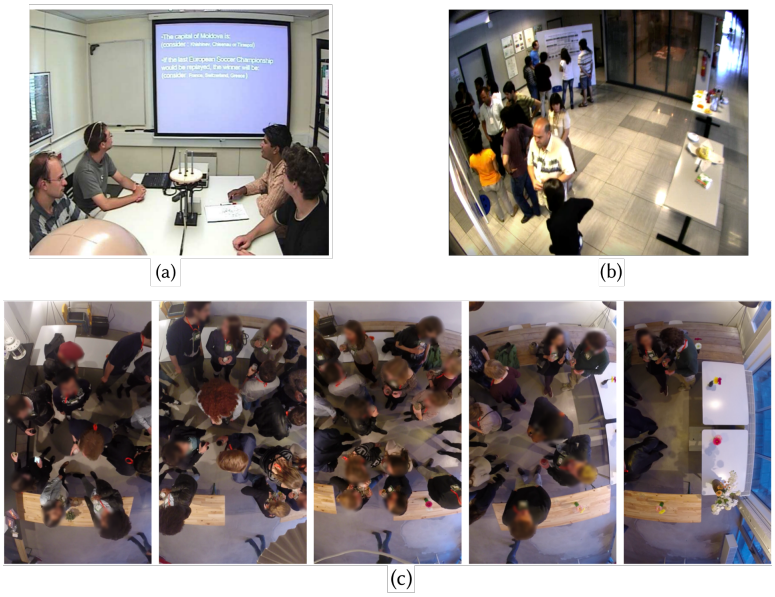


Fig. 1. Examples of types of scenes where people interact with each other in groups. (a)[5] and (b)[3] show a meeting and poster session in which participants have shared targets of attention (screen, poster board, etc.). In (c)[10], the unstructuredness makes studying the social interactions hard.

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 [5] (Figure 1(a)), OpenFace [7], etc., among other methods that use audio data [1, 56]). For our setting (Figure 1(b) and Figure 1(c)), adopting a wired connection for direct 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.,[52, 58, 60])) 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 [52] 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 [52, 68]. The Panoptic studio [29] 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 [3, 59] 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 [10] and [36]. 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., [3, 52, 58–60]) 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* [32], 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 [10][1].

In Section 2, we review the related works on head orientation estimation. Section 3 presents the details of our approach, methodology, and implementation. Section 4 discusses the dataset and our design choice based on it. Section 5 presents the relevant results and analyses. We close by discussing future work and conclusion in Sections 6 and 7.

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 2.1, we summarize related works in orientation measurement and estimation using wearable sensors. In Section 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.

2.1 Wearable Sensor Based Orientation Measurement and Estimation

IMUs containing accelerometer, gyroscope, and magnetometer can be used to obtain orientation estimations. Kok et al. [34] 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 [51]. Many motion capture systems feature large number of IMUs attached to custom made suits to capture positions and orientations in everyday surroundings [62]. However, these approaches do not align with the reconstruction using video and do not address the drifting issue [63]. To compensate for the drift, multiple cameras and IMUs are combined to estimate orientations [47, 64]. 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 [63, 64]. 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 [5] 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 [40], have been proposed. However, their study has shown that great variability exists in how people wear the sensor in the ear

[1] Available upon requesting access from http://matchmakers.ewi.tudelft.nl/matchnmingle/pmwiki/

(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) [18, 23], 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 [41]. 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 [31]. Such practice would violate the ecological validity [50] of the observations of human behaviors. On the other hand, the use of wearable badges around the neck is less intrusive [65] and could be used in settings where devices such as Google Glass or cell phones are considered impolite or forbidden.

2.2 Room Mounted Camera-Based Head Orientation Estimation

Image-based head orientation estimation is represented by a large body of literature [8, 26, 38, 45, 54, 66–70]. 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 [13, 24] 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. [52] 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° [3, 59] 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. [68] 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. [48] 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 [3, 52, 58, 60]. 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. [42] 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 [5, 39], 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.

3 METHODOLOGY

3.1 Approach

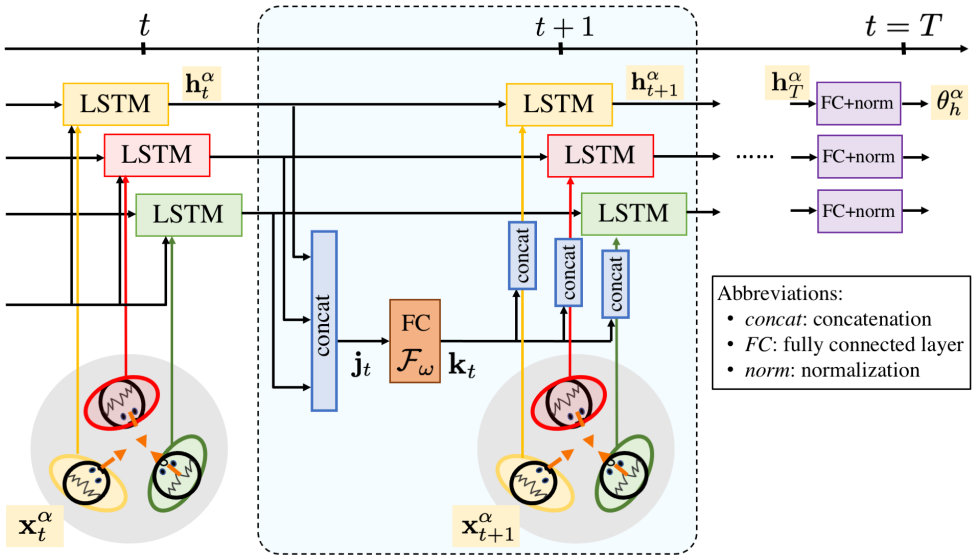


Fig. 2. Graphical illustration of our approach for head orientation estimation in the group context.

For a given group of size G, let $\boldsymbol{x}_t^\alpha \in \mathbb{R}^N$ denote the feature vector for member $\alpha \in \mathcal{G}$ at sequence step $t \in \{1, \ldots, T\}$, where $\mathcal{G} = \{1, \ldots, G\}$ is the set of members in the group, T is the sequence length, and N is the number of features. The feature vector $\boldsymbol{x}_t^\alpha$ 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. Let $\boldsymbol{h}_t^\alpha \in \mathbb{R}^H$ denote the hidden states associated with member α at sequence step t, where H is the number of hidden states (chosen as a hyperparamter). The hidden states at $t = 1$ are initialized as $\boldsymbol{h}_1^\alpha = \mathbf{0}$. 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 $\boldsymbol{j}_t \in \mathbb{R}^{G \times H}$ as

$$\boldsymbol{j}_t = [\boldsymbol{h}_t^1; \boldsymbol{h}_t^2; \ldots; \boldsymbol{h}_t^G], \tag{1}$$

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

$$\boldsymbol{k}_t = \mathrm{ReLU}(\mathcal{F}_\omega(\boldsymbol{j}_t)), \tag{2}$$

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 $\boldsymbol{k}_t$. Thereafter, for each member α, the respective hidden state $\boldsymbol{h}_t^\alpha$ as well as the concatenation of $\boldsymbol{x}_t^\alpha$ (individual member's input) and $\boldsymbol{k}_t$ (group-level context) are passed to an LSTM cell $\mathcal{L}_\tau$ (parameterized by the set τ) to obtain the output hidden states

$$\boldsymbol{h}_{t+1}^\alpha = \mathcal{L}_\tau([\boldsymbol{x}_{t+1}^\alpha; \boldsymbol{k}_t], \boldsymbol{h}_t^\alpha). \tag{3}$$

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

$$\begin{aligned}
\boldsymbol{f}_{t+1}^\alpha &= \sigma\left(\mathcal{W}_{\xi_f}\left([\boldsymbol{x}_{t+1}^\alpha; \boldsymbol{k}_t; \boldsymbol{h}_t^\alpha]\right)\right) && \text{(forget gate's activation vector)}\\
\boldsymbol{i}_{t+1}^\alpha &= \sigma\left(\mathcal{W}_{\xi_i}\left([\boldsymbol{x}_{t+1}^\alpha; \boldsymbol{k}_t; \boldsymbol{h}_t^\alpha]\right)\right) && \text{(input gate's activation vector)}\\
\boldsymbol{o}_{t+1}^\alpha &= \sigma\left(\mathcal{W}_{\xi_o}\left([\boldsymbol{x}_{t+1}^\alpha; \boldsymbol{k}_t; \boldsymbol{h}_t^\alpha]\right)\right) && \text{(output gate's activation vector)}\\
\tilde{\boldsymbol{c}}_{t+1}^\alpha &= \tanh\left(\mathcal{W}_{\xi_c}\left([\boldsymbol{x}_{t+1}^\alpha; \boldsymbol{k}_t; \boldsymbol{h}_t^\alpha]\right)\right) && \text{(cell input activation vector)}\\
\boldsymbol{c}_{t+1}^\alpha &= \boldsymbol{f}_{t+1}^\alpha \odot \boldsymbol{c}_t^\alpha + \boldsymbol{i}_{t+1}^\alpha \odot \tilde{\boldsymbol{c}}_{t+1}^\alpha && \text{(cell state vector)}\\
\boldsymbol{h}_{t+1}^\alpha &= \boldsymbol{o}_{t+1}^\alpha \odot \tanh\left(\boldsymbol{c}_{t+1}^\alpha\right) && \text{(output hidden state vector),}
\end{aligned} \tag{4}$$

where σ is sigmoid activation, $\odot$ denotes the Hadamard product, and $\boldsymbol{c}_t^\alpha$ and $\boldsymbol{c}_{t+1}^\alpha$ 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) upto $t = T$, the hidden states $\boldsymbol{h}_T^\alpha$ are passed through another linear layer $\mathcal{K}_\kappa$ (parameterized by κ),

$$\boldsymbol{q}^\alpha = \mathcal{K}_\kappa(\boldsymbol{h}_T^\alpha), \tag{5}$$

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

$$\boldsymbol{y}^\alpha = \frac{\boldsymbol{q}^\alpha}{\|\boldsymbol{q}^\alpha\|}. \tag{6}$$

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, $\boldsymbol{y}^\alpha = (y_1, y_2)^\intercal$ may be interpreted as cosine and sine of the head orientation prediction θ_h^α. To this end, we obtain θ_h^α from $\boldsymbol{y}^\alpha$ as

$$\theta_h^\alpha = \text{atan2}(y_2, y_1), \tag{7}$$

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 [9, 48] summed up over all the members in $\mathcal{G}$

$$\ell = \sum_{\alpha=1}^{G} \left(1 - \boldsymbol{y}^\alpha \cdot \left(\cos\theta_{h,GT}^\alpha, \sin\theta_{h,GT}^\alpha\right)^\intercal\right), \tag{8}$$

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

$$\omega^*, \tau^*, \kappa^* = \underset{\omega,\tau,\kappa}{\text{argmin}} \sum_{i=1}^{n} \ell_i, \tag{9}$$

where ℓ_i is the loss associated for the ith sequence in the dataset. Note that the group size (G) may vary within the dataset.

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[2] 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 [32].

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 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 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 [10]). 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 $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\Delta_j^2}$ over n test samples, where we define the angular difference Δ as

$$\Delta(\theta_1, \theta_2) = \min\left(|\theta_1 - \theta_2|, 2\pi - |\theta_1 - \theta_2|\right), \quad \text{with} \quad \theta_1, \theta_2 \in (-\pi, \pi], \tag{10}$$

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

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.

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

The rule-based method of head orientation estimation is inspired by previous works [5, 43, 44, 53] 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 3. Each colored area indicates the area of influence of an individual. The hexagon in Figure 3(b) delineates an interaction space containing three individuals interacting with one another, as an example. Figures 3(a) and 3(c) represent the frame-based and sequence-based method, respectively, and the estimation only concerns the bottom individual (denoted in pink). Figure 3(a) shows that the frame-based method only considers inputs at a single time step, whereas Figure 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 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 3(d)).

4 HEAD ORIENTATIONS IN COMPLEX SOCIAL SCENES: A CASE STUDY

4.1 Dataset

We develop and test our model on the MatchNMingle dataset [10] 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

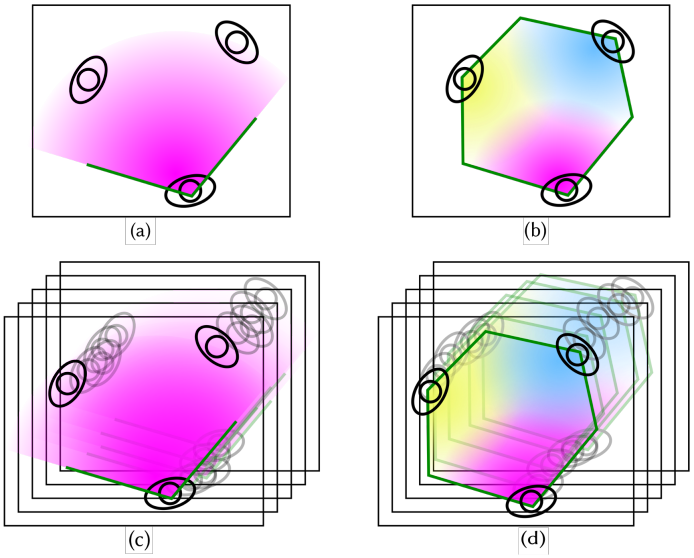


Fig. 3. Conceptual visualization of the motivation behind different methods. (a) The inputs and output are designed from an individual's perspective in the group. The frame-based method is designed for this scenario. (b) The inputs and outputs are designed from a group perspective. (c) A sequential version 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.

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 [21]. 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 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 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 5.

The group size distribution of all the three days in the MatchNMingle data is shown in Figure 4. Smaller groups are more common than larger groups. Figure 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.

Table 1. Comparison of different datasets available/used for head orientation estimation. Here, annnotations 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 [5]	seated meeting	4; ×8 meetings	10; ×8 meetings	Video	N/A	FOB† sensor
TownCentre[8]	pedestrian	2200	22	Video	N/A	Human
CAVIAR[16]	pedestrian (scripted)	40	1; ×17 segments	Video	25	Human
CocktailParty [52] ‡	FCG	6	30	Video	N/A	Automatic
Coffeebreak[15]	FCG	14	2; ×2 segments	Video	N/A	Automatic
SALSA-Cocktail Party[2]	FCG	18	25	Video, Wearables	1/3	Human
MatchNMingle[10]	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 [52, 60], 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.

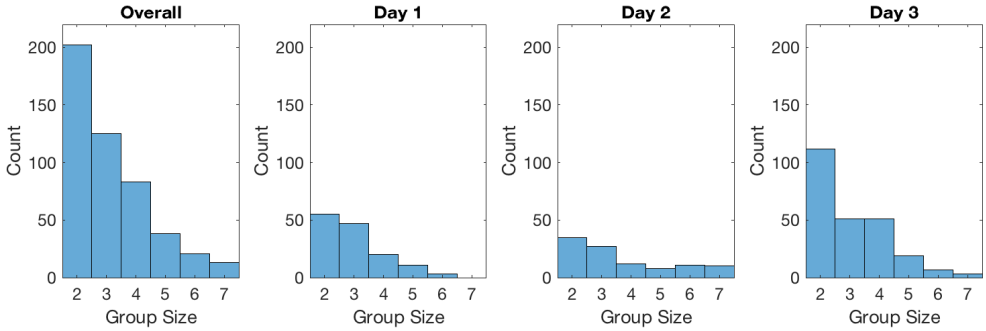


Fig. 4. Distribution of group sizes in the MatchNMingle dataset.

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 $17 \pm 5°$.

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

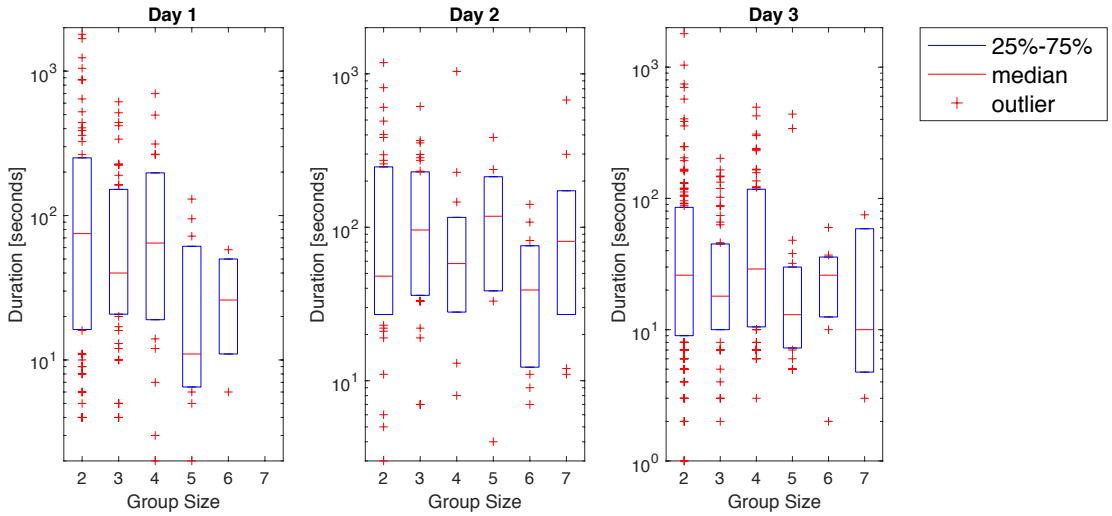


Fig. 5. Distribution of group durations in the MatchNMingle dataset.

MatchNMingle dataset in Figure 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 [2, 52, 58–60]). 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 5.

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 $T = 10$.

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. In Section 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 5.2-5.3. In Section 5.4, we compare our proposed method which is based on regression to an 8-class classification setting. In Section 5.5, we assess the model sensitivity with respect to body orientations. And lastly in Section 5.6, we show how our model handles dynamic moments in an interaction better than the purely socially-motivated rule-based method.

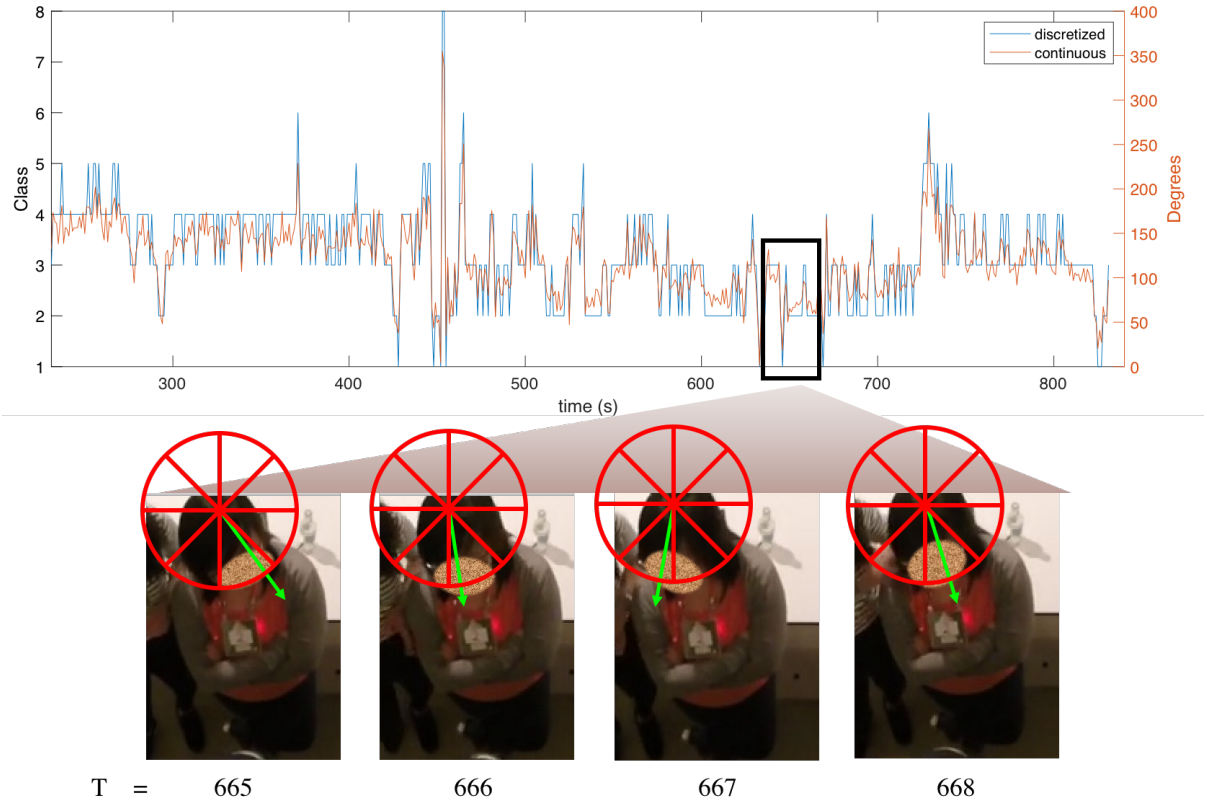


Fig. 6. Illustrative example of how small but noticeable head turns could result in the same class bin, resulting in a loss of fine-grained information. From T= 665 to T=666, the head turn with a relatively large magnitude resulted in the same bin. However, from T=666 to T=667, a head with a smaller magnitude resulted in different bins.

We perform all the experiments following the three-fold cross-validation scheme introduced Section 3.2 for each day in the dataset. All methods are evaluated with the same cross-validation splits. For the generalization evaluations in Section 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.

5.1 Model Comparison

In Table 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 $p < 0.01$.

Table 2. Summary of head orientation estimation performance of the proposed method and the baseline methods (introduced in Section 3.3). All methods are trained and evaluated with the same cross-validation splits.

Method	Type	Mean RMSE in θ_h (std. dev.) [°]		
		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)

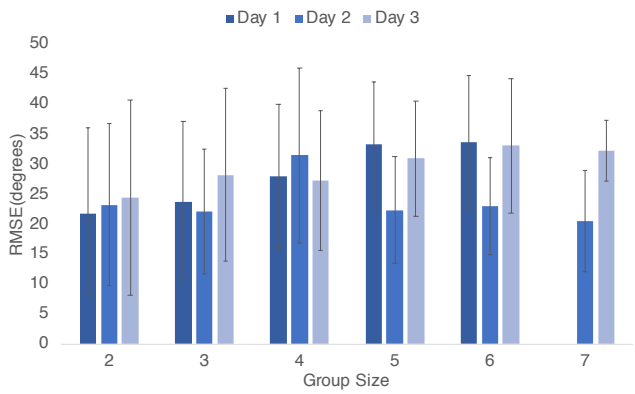


Fig. 7. Group-size-wise performance of the proposed group-based method in head orientation estimation. Note that there are no groups with seven members in the Day 1.

We further report the group-size-wise performance of the proposed method in Figure 7. Dyads typically have the lowest errors while larger groups have higher errors. This aligns with the intuition and corroborates previous observations [20] 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 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 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 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 4), with the difference in performance being small compared to the same for the rule-based method.

5.2 Generalization to Unseen Data

5.2.1 Cross-Day Generalization within the MatchNMingle Dataset. The three days for which the MatchNMingle data are available have different distributions of group sizes (Figure 4) as well as qualitative differences that include demographics, personality, acquaintance level, etc., of the participants. To this end, we assess the generalization

Table 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.) [°]		
	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)

Table 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.) [°]		
	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 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

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

5.2.2 Cross-Dataset Generalization to SALSA-Cocktail-Party Dataset. As shown in Table 1, the SALSA-Cocktail-Party dataset [2] 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 $22.5 \pm 14.0°$, which implies that the proposed model is capable of generalization to different datasets.

Table 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.) [°]						
	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 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.) [°]	
	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)

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 6). Body orientation contributes the most to the performance, which corroborates previous observations in head orientation estimation in pedestrian [14] and social settings like poster session [3]. 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)).

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 classification task which is the more common approach [3, 12, 52, 59]. 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 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.

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

Table 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.) [°]	
	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)

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 [30] 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.

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.

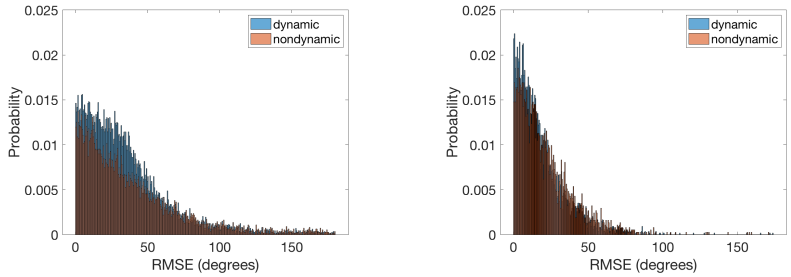
We propose a scheme to categorize samples as dynamic or nondynamic depending on speaking status switches to further showcase the efficacy of our proposed method. For a given time t, a fixed interval of window w is defined such that the speaking status sequence $[t - w, t]$ is considered. For the speaking 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 8. Figure 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 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.

6 DISCUSSION AND FUTURE WORK

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 [33, 36]. Gedik and Hung [19] 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).



(a) rule-based (b) group-based

Fig. 8. RMSE distribution of Day 1 validation results using the rule-based and the proposed group-based approach.

- **Body orientations:** Wearable light tags introduced by Montanari et al. [41] 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 [22], bluetooth beacons [17], radio frequency identification tags [55], 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 [20] 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.

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 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 [4–6, 43, 44]) which leverage the prior knowledge are worth considering if data resources are scarce and if more model interpretability is preferred.

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 [37]. 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.

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 [20] 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.

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.

ACKNOWLEDGMENTS

This research was partially funded by the Netherlands Organization for Scientific Research (NWO) under the MINGLE (Modelling Social Group Dynamics and Interaction Quality in Complex Scenes using Multi-Sensor Analysis of Non-Verbal Behaviour) project number 639.022.606. We thank Felix van Doorn and Lu Liu for annotating a subset of F-formation membership in the data.

REFERENCES

[1] Alberto Abad, Carlos Segura, Climent Nadeu, and Javier Hernando. 2007. Audio-based approaches to head orientation estimation in a smart-room. In *Eighth Annual Conference of the International Speech Communication Association*.

[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] Xavier Alameda-Pineda, Yan Yan, Elisa Ricci, Oswald Lanz, and Nicu Sebe. 2015. Analyzing Free-standing Conversational Groups: A Multimodal Approach. In *Proceedings of the 23rd ACM International Conference on Multimedia (MM '15)*. ACM, New York, NY, USA, 5–14. https://doi.org/10.1145/2733373.2806238

[4] Sileye O. Ba, Hayley Hung, and Jean-Marc Odobez. 2009. 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)*. IEEE Press, 1424–1427.

[5] Sileye O Ba and Jean-Marc Odobez. 2009. Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 1 (2009), 16–33.

[6] Sileye O. Ba and Jean-Marc Odobez. 2011. Multiperson Visual Focus of Attention from Head Pose and Meeting Contextual Cues. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 1 (Jan. 2011), 101–116. https://doi.org/10.1109/TPAMI.2010.69

[7] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[8] Ben Benfold and Ian Reid. 2008. Colour Invariant Head Pose Classification in Low Resolution Video. In *Proceedings of the 19th British Machine Vision Conference*.

[9] Lucas Beyer, Alexander Hermans, and Bastian Leibe. 2015. Biternion nets: Continuous head pose regression from discrete training labels. In *German Conference on Pattern Recognition*. Springer, 157–168.

[10] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. v. d. Meij, and H. 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* (2018), 1–1. https://doi.org/10.1109/TAFFC.2018.2848914

[11] Tanya L Chartrand and John A Bargh. 1999. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology* 76, 6 (1999), 893.

[12] Cheng Chen, Alexandre Heili, and Jean-Marc Odobez. 2011. Combined estimation of location and body pose in surveillance video. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*. IEEE, 5–10.

[13] Cheng Chen, Alexandre Heili, and Jean-Marc Odobez. 2011. A joint estimation of head and body orientation cues in surveillance video. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 860–867.

[14] Cheng Chen and Jean-Marc Odobez. 2012. 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*. IEEE, 1544–1551.

[15] 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. 4.

[16] James L Crowley, Patrick Reignier, and Sebastien Pesnel. 2004. Context Aware Vision using Image-based Active Recognition. (2004).

[17] Ramsey Faragher and Robert Harle. 2015. Location fingerprinting with bluetooth low energy beacons. *IEEE journal on Selected Areas in Communications* 33, 11 (2015), 2418–2428.

[18] Andrea Ferlini, Alessandro Montanari, Cecilia Mascolo, and Robert Harle. 2019. Head Motion Tracking Through in-Ear Wearables. (2019).

[19] Ekin Gedik and Hayley Hung. 2017. Personalised models for speech detection from body movements using transductive parameter transfer. *Personal and Ubiquitous Computing* 21, 4 (2017), 723–737.

[20] 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), 163.

[21] Gerard E Grossman, R John Leigh, Larry A Abel, Douglas J Lanska, and SE Thurston. 1988. Frequency and velocity of rotational head perturbations during locomotion. *Experimental brain research* 70, 3 (1988), 470–476.

[22] Dominik Gusenbauer, Carsten Isert, and Jens Krösche. 2010. Self-contained indoor positioning on off-the-shelf mobile devices. In *2010 International Conference on Indoor Positioning and Indoor Navigation*. IEEE, 1–9.

[23] Taku Hachisu, Yadong Pan, Soichiro Matsuda, Baptiste Bourreau, and Kenji Suzuki. 2018. FaceLooks: A Smart Headband for Signaling Face-to-Face Behavior. *Sensors* 18, 7 (2018), 2066.

[24] Irtiza Hasan, Francesco Setti, Theodore Tsesmelis, Alessio Del Bue, Fabio Galasso, and Marco Cristani. 2018. MX-LSTM: mixing tracklets and vislets to jointly forecast trajectories and head poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6067–6076.

[25] Jari K Hietanen. 2002. Social attention orienting integrates visual information from head and body orientation. *Psychological Research* 66, 3 (2002), 174–179.

[26] Thanarat Horprasert, Yaser Yacoob, and Larry S Davis. 1996. Computing 3-d head orientation from a monocular image sequence. In *Proceedings of the second international conference on automatic face and gesture recognition*. IEEE, 242–247.

[27] Hayley Hung, Ekin Gedik, and Laura Cabrera Quiros. 2019. Complex conversational scene analysis using wearable sensors. In *Multimodal Behavior Analysis in the Wild*. Elsevier, 225–245.

[28] Dušan Jan and David R Traum. 2007. Dynamic movement and positioning of embodied agents in multiparty conversations. In *Proceedings of the Workshop on Embodied Language Processing*. Association for Computational Linguistics, 59–66.

[29] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. 2017. Panoptic studio: A massively multiview system for social interaction capture. *IEEE transactions on pattern analysis and machine intelligence* 41, 1 (2017), 190–204.

[30] I Kåsa. 1976. A circle fitting procedure and its error analysis. *IEEE Transactions on instrumentation and measurement* 1 (1976), 8–14.

[31] Norene Kelly. 2017. All the world's a stage: what makes a wearable socially acceptable. *interactions* 24, 6 (2017), 56–60.
[32] Adam Kendon. 1970. Movement coordination in social interaction: Some examples described. *Acta psychologica* 32 (1970), 101–125.
[33] Taemie Kim, Agnes Chang, and Alex Sandy Pentland. 2007. Enhancing organizational communication using sociometric badges. In *Proceedings of the 11th International Symposium on Wearable Computers (Submitted)*.
[34] Manon Kok, Jeroen D Hol, and Thomas B Schön. 2017. Using inertial sensors for position and orientation estimation. *arXiv preprint arXiv:1704.06053* (2017).
[35] Stephen RH Langton, Roger J Watt, and Vicki Bruce. 2000. Do the eyes have it? Cues to the direction of social attention. *Trends in cognitive sciences* 4, 2 (2000), 50–59.
[36] Oren Lederman, Dan Calacci, Angus MacMullen, Daniel C Fehder, Fiona E Murray, and Alex'Sandy' Pentland. 2017. Open badges: A low-cost toolkit for measuring team communication and dynamics. *arXiv preprint arXiv:1710.01842* (2017).
[37] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. 2018. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730* (2018).
[38] Yang Lu, Shujuan Yi, Nan Hou, Jingfu Zhu, and Tiemin Ma. 2016. Deep neural networks for head pose classification. In *Intelligent Control and Automation (WCICA), 2016 12th World Congress on*. IEEE, 2787–2790.
[39] Benoît Massé, Silèye O. Ba, and Radu Horaud. 2017. Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction. *CoRR* abs/1703.04727 (2017).
[40] Chulhong Min, Akhil Mathur, Alessandro Montanari, and Fahim Kawsar. 2019. An early characterisation of wearing variability on motion signals for wearables. In *Proceedings of the 23rd International Symposium on Wearable Computers*. 166–168.
[41] Alessandro Montanari, Zhao Tian, Elena Francu, Benjamin Lucas, Brian Jones, Xia Zhou, and Cecilia Mascolo. 2018. Measuring interaction proxemics with wearable light tags. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 25.
[42] Kazuhiro Otsuka, Keisuke Kasuga, and Martina Köhler. 2018. Estimating Visual Focus of Attention in Multiparty Meetings using Deep Convolutional Neural Networks. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 191–199.
[43] Kazuhiro Otsuka, Yoshinao Takemae, and Junji Yamato. 2005. 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*. 191–198.
[44] Kazuhiro Otsuka, Junji Yamato, Yoshinao Takemae, and Hiroshi Murase. 2006. Conversation scene analysis with dynamic bayesian network basedon visual head tracking. In *2006 IEEE International Conference on Multimedia and Expo*. IEEE, 949–952.
[45] Chavdar Papazov, Tim K Marks, and Michael Jones. 2015. 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*. 4722–4730.
[46] Alex Pentland. 2014. *Social physics: How good ideas spread-the lessons from a new science*. Penguin.
[47] Gerard Pons-Moll, Andreas Baak, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bodo Rosenhahn. 2010. Multisensor-fusion for 3d full-body human motion capture. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 663–670.
[48] Sergey Prokudin, Peter Gehler, and Sebastian Nowozin. 2018. Deep directional statistics: Pose estimation with uncertainty quantification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 534–551.
[49] Chirag Raman and Hayley Hung. 2019. Towards automatic estimation of conversation floors within F-formations. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 175–181.
[50] Arran T Reader and Nicholas P Holmes. 2016. Examining ecological validity in social interaction: problems of visual fidelity, gaze, and social potential. *Culture and Brain* 4, 2 (2016), 134–146.
[51] Valérie Renaudin, Muhammad Haris Afzal, and Gérard Lachapelle. 2010. New method for magnetometers based orientation estimation. In *IEEE/ION Position, Location and Navigation Symposium*. IEEE, 348–356.
[52] Elisa Ricci, Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulo, Narendra Ahuja, and Oswald Lanz. 2015. 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*. 4660–4668.
[53] Rutger Rienks, Ronald Poppe, and Dirk Heylen. 2005. Differences in head orientation between speakers and listeners in multi-party conversations. *International Journal HCS* (2005).
[54] Nataniel Ruiz, Eunji Chong, and James M Rehg. 2018. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2074–2083.
[55] Samer S Saab and Zahi S Nakad. 2010. A standalone RFID indoor positioning system using passive tags. *IEEE Transactions on Industrial Electronics* 58, 5 (2010), 1961–1970.
[56] Carlos Segura, Cristian Canton-Ferrer, Alberto Abad, Josep R Casas, and Javier Hernando. 2007. Multimodal head orientation towards attention tracking in smartrooms. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 2. IEEE, II–681.

[57] 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.

[58] Ramanathan Subramanian, Yan Yan, Jacopo Staiano, Oswald Lanz, and Nicu Sebe. 2013. 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)*. ACM, New York, NY, USA, 3–10. https://doi.org/10.1145/2522848.2522862

[59] Stephanie Tan, David MJ Tax, and Hayley Hung. 2018. 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*. 1–8.

[60] Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulò, Narendra Ahuja, Oswald Lanz, and Elisa Ricci. 2018. Joint Estimation of Human Pose and Conversational Groups from Social Scenes. *International Journal of Computer Vision* 126, 2-4 (2018), 410–429.

[61] Troy AW Visser and Ashton Roberts. 2018. Automaticity of social cues: The influence of limiting cognitive resources on head orientation cueing. *Scientific reports* 8, 1 (2018), 10288.

[62] Daniel Vlasic, Rolf Adelsberger, Giovanni Vannucci, John Barnwell, Markus Gross, Wojciech Matusik, and Jovan Popović. 2007. Practical motion capture in everyday surroundings. *ACM transactions on graphics (TOG)* 26, 3 (2007), 35–es.

[63] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 601–617.

[64] Timo Von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. 2016. Human pose estimation from video and imus. *IEEE transactions on pattern analysis and machine intelligence* 38, 8 (2016), 1533–1547.

[65] Benjamin N Waber, Daniel Olguin Olguin, Taemie Kim, and Alex Pentland. 2008. Understanding organizational behavior with wearable sensing technology. *Available at SSRN 1263992* (2008).

[66] Ying Wu and Kentaro Toyama. 2000. Wide-range, person-and illumination-insensitive head orientation estimation. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. IEEE, 183–188.

[67] Jing Xiao, Tsuyoshi Moriyama, Takeo Kanade, and Jeffrey F Cohn. 2003. Robust full-motion recovery of head by dynamic templates and re-registration techniques. *International Journal of Imaging Systems and Technology* 13, 1 (2003), 85–94.

[68] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Gaowen Liu, Oswald Lanz, and Nicu Sebe. 2015. A multi-task learning framework for head pose estimation under target motion. *IEEE transactions on pattern analysis and machine intelligence* 38, 6 (2015), 1070–1083.

[69] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. 2019. 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*. 1087–1096.

[70] Xiangxin Zhu and Deva Ramanan. 2012. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2879–2886.