Conversation Group Detection With Spatio-Temporal Context

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

DOI
10.1145/3536221.3556611

Publication date
2022

Document Version
Final published version

Published in
ICMI 2022 - Proceedings of the 2022 International Conference on Multimodal Interaction

Conversation Group Detection With Spatio-Temporal Context

Stephanie Tan
Delft University of Technology

David M.J. Tax
Delft University of Technology

Hayley Hung
Delft University of Technology

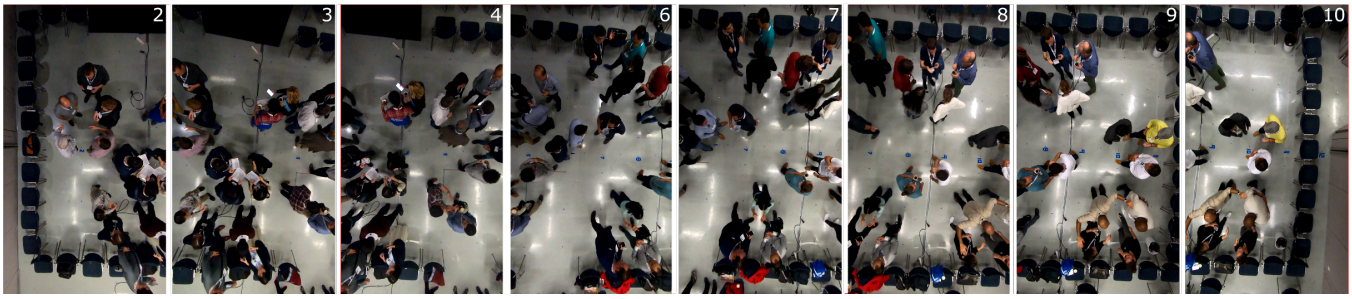

Figure 1: Snapshot of a typical social interaction area from the Conflab [34] dataset. Faces blurred for privacy.

ABSTRACT

In this work, we propose an approach for detecting conversation groups in social scenarios like cocktail parties and networking events, from overhead camera recordings. We posit the detection of conversation groups as a learning problem that could benefit from leveraging the spatial context of the surroundings, and the inherent temporal context in interpersonal dynamics which is reflected in the temporal dynamics in human behavior signals, an aspect that has not been addressed in recent prior works. This motivates our approach which consists of a dynamic LSTM-based deep learning model that predicts continuous pairwise affinity values indicating how likely two people are in the same conversation group. These affinity values are also continuous in time, since relationships and group membership do not occur instantaneously, even though the ground truths of group membership are binary. Using the predicted affinity values, we apply a graph clustering method based on Dominant Set extraction to identify the conversation groups. We benchmark the proposed method against established methods on multiple social interaction datasets. Our results showed that the proposed method improves group detection performance in data that has more temporal granularity in conversation group labels. Additionally, we provide an analysis in the predicted affinity values in relation to the conversation group detection. Finally, we demonstrate the usability of the predicted affinity values in a forecasting framework to predict group membership for a given forecast horizon.

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9390-4/22/11.
https://doi.org/10.1145/3536221.3556611

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; *Scene understanding*; • **Human-centered computing** → *Human computer interaction (HCI).*

ACM Reference Format:
Stephanie Tan, David M.J. Tax, and Hayley Hung. 2022. Conversation Group Detection With Spatio-Temporal Context. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3536221.3556611

1 INTRODUCTION

The automatic detection of conversation groups is an interesting problem for applications ranging from social surveillance [12, 23, 41] to social robotics [7, 38, 48]. In social settings, such as cocktail parties or professional networking events, the floor for interactions at the venue consists of multiple conversation groups that dynamically adapt to the ebb and flow of the underlying human behaviors that determine the social interactions. Characterizing the interpersonal relationships (the level of spontaneous affinity between people) that foster individuals to freely congregate to form a focused encounter and exchange information could help us understand more about interaction experience and quality [27, 32]. However, due to the complexity and subtleties in human social dynamics in changing environments which are also context-specific, automatically detecting conversation groups in social scenes is an ongoing and challenging research topic.

For this paper, we focus on identifying conversation groups (more specifically, free-standing conversation groups (FCGs) [41]) in a social scene where individuals physically come together to interact with each other. In these settings they organize themselves into groups that define the physical partitions of who is interacting with whom. Figure 1 shows a representative scene for the types of social scenarios that we are interested in from an overhead viewpoint. People's use of physical space is known as *proxemics* [20].The conversation groups can be of varying duration, size, and spatial arrangement. In practice, these conversation groups have often been conceptually formalized as *facing-formation*

(F-formations) [3, 23, 41]. As introduced and defined by Kendon [26], an F-formation is formed when two or more people arrange themselves onto a convex envelope to enclose an overlap of their interaction transactional segments (i.e., space in front of them where sight and hearing are most effective [9]). The interactants have equal and exclusive access to this overlap (i.e., *o-space*).

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

A social scene such as Figure 1 may be represented by an interaction graph with nodes representing the individuals, and edges representing the relationship between two individuals. Conversation groups can be deduced from the information in this interaction graph, where individuals in the same conversation have greater edge weights (affinity) with one another, and vice versa. Each individuals have behavioral cues such as positions, head/body orientations, etc. that are shaped by the surroundings, and also change over time. Indeed, proximity with directional information could already be indicative of group membership if one simply considers 'who is standing with whom'. However, in social scenes such as Figure 1, factors such as crowdedness and furniture layout define the spatial context that affects individual cues that determines conversation group membership. The temporal context in the behavioral cues plays a role in the interactions which are dynamic in nature. One type of social dynamics that could be captured by movement cues (e.g., change in positions and orientations) is synchrony and mimicry patterns which are known to be important driver for affiliation and interpersonal rapport [17, 22, 42]. In terms of conversation dynamics, since multiple conversation floors are possible within one F-formation [33], head orientations that change during turn-taking or other types of conversation dynamics are especially relevant to conversation schisms. The schisming phenomenon may lead to two or more distinct conversation groups [13] and hence possibly form new F-formations. Therefore, in this paper, we argue that even though F-formation in its original definition is a static concept, it is important to take into account the temporal context of the behavioral cues influencing the groups.

The second fold of the challenge is the potential problem in how ground truths of group membership are defined from pairwise affinities. Past approaches of automatic detection of conversation groups have identified affinity and group memberships as binary and operationalized the task based on this design choice, albeit Zhang and Hung have investigated the subjectivity of annotating groups [51]. Although the assumption of the binary group membership which existing methods and ours hinge on, is valid to the extent of how they are reflected in the ground truth annotations, interpersonal relationships are not binary in reality. The temporally evolving affinity between two individuals, does not change from zero to one, or one to zero, instantaneously [18, 19, 46]. Social interactions have a rite of passage, from greeting to leaving [4]. We aim to understand how the affinity scores change in time and encapsulate the changing dynamics of the behavioral cues, and how they affect group detection, which are not apparent in hard assignments of group memberships. Our paper takes a step towards this direction which has not been the focus in previous works.

Following prior works [43, 45], our approach to conversation group detection consists of two stages: (1) we first estimate continuous pairwise affinities between all individuals in a social scene, and (2) we use an existing framework to cluster the individuals by leveraging a graph clustering based on Dominant Set (DS) to identify conversation groups [23]. In order to account for the temporal context in the behavioral cues, we introduce a deep learning based Long Short-Term Memory (LSTM) network to predict pairwise affinity scores that determine the F-formation membership (annotated as ground truth of conversation groups in existing datasets). The inputs to our network are temporally aligned sequential inputs (including positions, and head and/or body orientation); the output of our network is the pairwise affinity value between one individual and *all* other members in the social interaction scene.

In contrast to existing works that output an affinity score of all pairs of individuals independently which scales quadratically with the number of people in the scene [43, 45], our design for person-wise output scales linearly and follows more intuitively from an egocentric application point of view (i.e. social robots). DS clustering is applied to the affinity matrix corresponding to each scene to detect the conversation groups. The clique formulation in DS is exploited to identify clusters in the interaction graph, which refines group detection since the affinity matrix (from estimated pairwise affinities) may not directly provide self-consistent and symmetric binary group memberships [21, 23, 43].

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

- We propose a novel LSTM-based affinity score model to approximate the likelihood of two people interacting in the same conversation group. The model includes a pooling module to account for the spatial context of social interactions, inspired by what [43, 45] captured in their models. Using the proposed model that leverages (temporal) sequential input features in addition to a pooling module, we simultaneously account for the spatio-temporal context that affects pairwise affinities that determine the conversation groups.
- We provide an analysis over the predicted affinity scores in conversation group detection, characterized by Area under (receiver-operating) Curve (AUC) scores, followed by qualitative examples showing the continuity of affinity scores. We also show a comparison of affinity score processing (asymmetric vs. symmetric) for application in DS and the group detection performance with respect to scene dynamics.
- We demonstrate the usability of the predicted affinity values via a novel forecasting framework for affinity score prediction based on Gaussian Process Regression (GPR). The

framework also provides inferential uncertainty quantification over the predictions of future conversation groups.

2 RELATED WORKS

Conversation group detection in situated interactions has been tackled by a variety of approaches stemming from different communities (computer vision [12, 36, 37, 41], human-computer interaction [5, 6, 10, 21, 48], etc.). This section discusses the representative works in this area. Conversation groups and the more formalized F-formation representation are analogous in terms of group detection in interaction scenes in past works.

Many previous works, especially from the vision community, use features such as location and head/body orientations for the task of group detection [3, 23, 37, 41, 47]. These quantities could be obtained automatically from vision data using multi-camera surveillance setups (typically elevated side-views).

Using these features, some methods for F-formation detection have been focused on optimization-based approaches to mathematically model the physical space. More concretely, a number of works [15, 40, 41] hypothesised that the o-space can be generated from a noisy representation of the instantaneous view frustrum obtained from the head pose. Heat maps generated from samples projected from each individual's view frustrum were then used to identify o-spaces. Members of the F-formation were then re-identified as belonging to a particular o-space based a pre-defined metric of closeness.

Another class of methods have formulated social scenes as an edge-weighted graph where each individual represents a node and the edge represents the pair-wise connection between individuals. These methods take the part of the F-formation definition related to equal mutual attention to synonymous with maximal cliques in edge weighted graphs. In early works, the pair-wise relationships were modeled using feature engineering based on location and orientation information [23, 47, 51, 52]. Aggregating these estimated pairwise affinity values, the affinity matrix serves as inputs to graph clustering based on Dominant Set using game-theoretic approaches [23, 47] to iteratively partition nodes to extract conversation groups. While the o-space is not explicitly modelled with these approaches, the maximal clique formulation implicitly models the o-space whilst also explicitly binding individuals to a specific group as part of the Dominant Set identification process. However, the representation of pairwise affinity, particularly when only location and orientation is used, forces a circular shape assumption to the F-formation that does not always happen in practice [51, 52]

To address this problem and enable more flexibility in modeling pairwise relationships given the surroundings, deep learning based approaches have been proposed. DANTE learns the affinity values by explicitly modeling dyadic and context interaction [43] by using relative positions and head/body orientations after preprocessing as inputs to the model. More recently, Thompson et al. [45] proposed a graph neural network (GNN) based approach that leverages the more general message-passing mechanism during training to predict affinities using raw signal data including absolute positions, accelerometer readings, and image. Similar to the preceding works [23, 47, 51], both of these deep learning based approaches also apply the learned affinities values inputs to DS graph clustering. Schmuck and Celiktutan [39] also proposed a GNN-based approach to predict interpersonal links, but as opposed to using DS graph clustering, a greedy agreement algorithm was applied to identify groups [21].

Departing from using visually obtained features such as locations and orientations of individuals, some works have taken advantage of features from other modalities that have shown to be helpful when estimating pairwise affinities. For example, [45] take advantage of a combination of motion based features and visually obtained features. Gedik and Hung have shown that it is possible to estimate groups purely based on motion features as phenomena such as body movement synchrony in interactions are indicative of pairwise relationships [16]. However, in communication with the authors, the predicted affinities need to be significantly improved for before DS clustering would yield reasonable performance. This highlights that the nature of the problem lies between the modelling social dynamics and proxemics (i.e. positions and orientations). Approaches proposed by the ubiquitous and pervasive computing communities have relied on Bluetooth Smart (BLE) to measure proximity values in terms of Received Signal Strength Indicator (RSSI) values which capture distance and orientation information to some extent, represented by [29]. In the case of [25], data from motion sensors (accelerometer and gyroscope) were also incorporated with proximity features for multimodal detection of groups using smart phones. Other custom sensors have been developed to measure proximity, relative orientation, motion, and/or a combination thereof (e.g., light tags [30], Rhythm badges [28] and the Midge [1]). These data also capture the useful information, such as direct measurement of closeness forming hypotheses of interactions already and the measurement of nuanced body motion, in determining conversation groups in social interactions, and methods developed based on these have the potential to scale more easily.

The proposed method differs from early works [23, 47, 51, 52] in F-formation detection as it is a deep learning method where pairwise affinities are learned. The main limitation is that the early works are based on feature engineering, which is less flexible than deep learning models. Compared to existing deep learning methods [39,43,45], the proposed method simultaneously models the spatio-temporal context in F-formation detection. We aim to capture the underlying evolving social dynamics by modeling the temporal dynamics in the input features, which has not been investigated before. Our method outperforms the most relevant deep learning baseline [43] for the ConfLab dataset, especially in scenes where there is high dynamicity.

3 APPROACH

The overview of the approach to conversation group detection is illustrated in Figure 2. Figure 2(a) represents an example image from an interaction scene. The individual attributes such as positions, head and body orientations encode spatial information of an individual with respect to the scene (labeled in Figure 2(b)).

Module (b) represents the core of our contribution, which is a novel deep learning neural network for pairwise affinity estimation, based on a joint Long Short-Term Memory (LSTM) network that simultaneously accounts for the temporal context of the input signals and spatial context in the scene with spatially-motivated

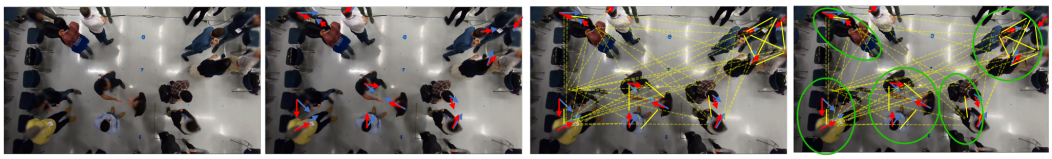

Figure 2: Visualization of the derivation of the interaction graph: (a): the given scene from an overhead camera; (b): each subject has person-wise features, such as positions, orientations, etc.; (c): the edge weights (affinity scores) between each pair of subjects are predicted via the proposed neural network model; and (d): using the DS [23], conversation groups are extracted as sets of subjects.

context pooling. In (c), the pairwise affinities are combined to a affinity matrix, and following previous approaches, Dominant Set was used to extract groups by iteratively identifying maximal cliques in edge-weighted graphs (module (d)). Our contribution focuses on the neural network architecture for affinity prediction, and assumes that the inputs are acquired and preprocessed upstream. We use the Dominant Set clustering method on graphs downstream of affinity prediction because it is state-of-the-art method for this use-case.

The details of the neural network architecture is described in Sec. 3.1, and the details of the dominant set method is described in in Sec. 3.2.

3.1 Affinity prediction

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

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

the details of which are discussed in Section 4.4.

Module (b) in Figure 3 demonstrates one recurrent step of the proposed model from time t to $t+1$. Let $\boldsymbol{h}_t^{j,i} \in \mathbb{R}^H$ denote the hidden representations associated with member i at sequence step t, where H is the dimension of hidden states (chosen as a hyperparameter). The hidden states at $t = 0$ are initialized as $\boldsymbol{h}_0^{j,i} = \mathbf{0}$. To capture the spatial context defined by all members in the interaction scene, we pool the hidden states of all the members $j \neq i$ as follows. To discount for the persons absent in the scene at time t, the hidden state are first processed through a masking layer (with element-wise multiplication) to obtain intermediate masked hidden states representation $\boldsymbol{g}_t^{j,i} = I_t^j \boldsymbol{h}_t^{j,i}$. Note that the masked representation reflects the presence of members j in the scene, which may be dynamically changing between different time steps. All masked representations $\boldsymbol{g}_t^{j,i}$ are processed through an *aggregate pooling* layer to obtain a scene level representation $\mathcal{K}_t^i$ given by

$$\mathcal{K}_t^i = \frac{\sum_{j \neq i} \boldsymbol{g}_t^{j,i}}{\sum_{j \neq i} I_t^j}. \tag{1}$$

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

$$\boldsymbol{o}_t^{j,i} = \lambda \mathcal{K}_t + (1 - \lambda) \boldsymbol{g}_t^{j,i}, \tag{2}$$

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

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

$$\boldsymbol{h}_{t+1}^{j,i} = \mathcal{L}_\tau([\boldsymbol{v}_{t+1}^{j,i}; \boldsymbol{o}_t^{j,i}], \boldsymbol{h}_t^{j,i}). \tag{3}$$

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

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

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

After the time roll out in each LSTM time step until $t = T$, the hidden states $\boldsymbol{h}_T^{j,i}$ are passed through a linear layer parameterized by set of weights $\boldsymbol{W}_m$ and biases $\boldsymbol{b}_m$ to obtain the final pairwise edge predictions with respect to member i. Subsequently, they

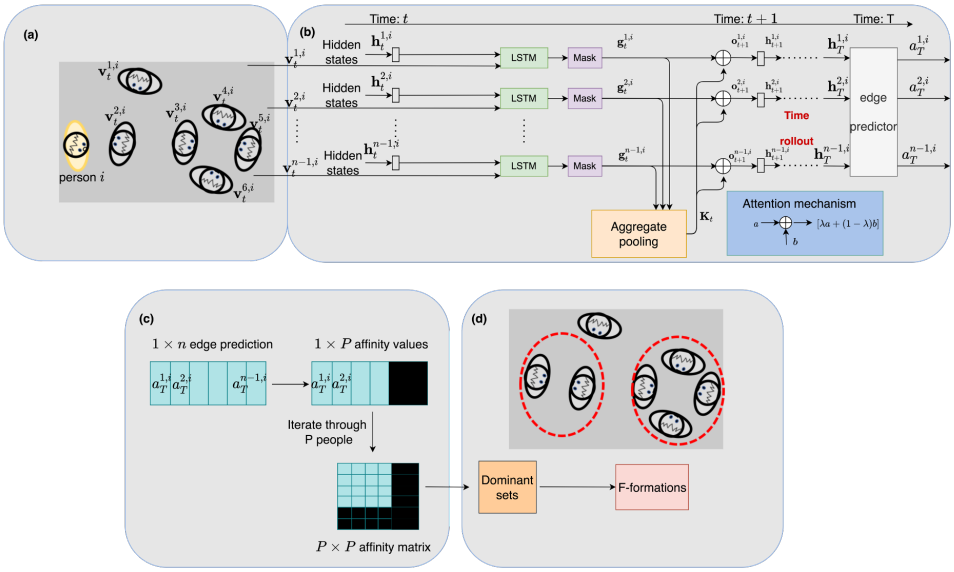

Figure 3: Visualization of the methodology. (a): a scene representation of individuals and preprocessed features with respect to person i; (b): a graphical representation of the time roll out from t to $t+1$ with the aggregate pooling layer along with the attention mechanism; (c) the filtering and aggregation of person-wise affinity output at $t = T$ into an affinity matrix; (d): application of DS clustering for group identification from the affinity matrix.

are passed through a sigmoid activation (denoted by σ) to obtain pairwise affinity $a^{j,i} \in [0, 1]$ as

$$a^{j,i} = \sigma(\boldsymbol{W}_m \boldsymbol{h}_T^{j,i} + \boldsymbol{b}_m), \tag{5}$$

where the values of 0 and 1 denote no and perfect pairwise affinity, respectively. The output of the model $a^{j,i}$ is strategically designed to be continuous, which lends naturally to a probability interpretation of pairwise interaction. We further motivate this choice, the performance evaluation, and the connection to downstream tasks such as conversation group forecasting in Section 4.3.

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

$$\ell = \sum_{S} \sum_{j \neq i} \left(a^{j,i} - a_{GT}^{j,i} \right)^2, \tag{6}$$

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

3.2 Dominant set clustering

As shown in Figure 3 (c), for each member i in the scene, the affinity prediction model predicts $n-1$ pairwise affinity values with respect to all other members j for all time steps irrespective of whether they are visible at that moment or not. To evaluate the group identification performance, we use the output at the last time step at $t = T$. After filtering with the indicator masks, a $P \times P$ affinity matrix for the scene in question is obtained, where P is the actual number of subjects at a particular scene.

After the predicted affinities are arranged into an affinity matrix, following prior approaches, [23, 43, 45, 47], the F-formations are extracted using Dominant Sets (DS) clustering. The resulting clusters representing F-formations could be of any size. The stopping criterion of the optimization formulation is either when the relative mutual affinity of internal nodes and external nodes of a dominant set do not satisfy the constraints, or when the mutual affinity of a group is lower than a chosen threshold. The second part of the stopping criterion enables improvement to detect singletons in the scene as it accounts for the global context (i.e., when there are only few people left after maximal clique extraction iterations, it is not likely that they are in the same group). We follow the implementation of F-formation clustering of [43]. For the theoretical background and more detailed reference to the application of Dominant Set Clustering for F-formation detection, please refer to [8, 23].

4 EXPERIMENTAL SETUP

4.1 Baseline methods

The baseline methods considered in this work are GTCG [47], GCFF [41], and DANTE [43]. GTCG and GCFF are both non deep learning based methods which rely on engineered position and orientation based features. GTCG models pairwise affinity values using distance between distributions over the plausible regions determined by the visual frustums, followed by a refined game theoretic approach for group extraction based on [23] and [11]. GCFF models the probability of individuals belonging to o-space centers (i.e. center of conversation groups), and uses a graph-cut approach in conjunction with constraint based on direct access to extract the groups.

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

approach takes advantage of image-based features and a rich collection of social action semantic labels, in addition to proxemics and body motion based features, for conversation group detection [45]. We omit comparison against this recent approach for the scope of the paper, since the focus of our paper is on modeling the social dynamics in conversation scene using temporal information rather than a thorough investigation of using different input modalities.

4.2 Datasets

To align with the closest state-of-the-art approaches, we compare our method on the same representative datasets including Cocktail Party [50] and SALSA [2]. We also report benchmark results on the recently released Conflab dataset [34], capturing professional networking social interactions in-the-wild.

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

Most of the existing datasets were collected to serve F-formation detection using visual information, i.e. using an elevated side-view. Bounding boxes and head/body orientations are acquired either through automated methods or manual annotations. For datasets that have a top-down view, positions and orientations are acquired through manual annotations because automated methods result in error prone inputs to subsequent models [34]. The overview of the datasets used is as follows:

-*Cocktail Party* [50]: contains 30 minute recordings of six people interacting with one another, captured by four elevated side view cameras in the corners of the space. Positions and head orientations of the subjects are obtained automatically using a particle filter-based body tracking method. The conversation groups were annotated at 1/5 Hz.

-*SALSA* [2]: contains 60 minute recordings of 18 people interacting with one another, captured by four elevated side view cameras in the corners of the space. Positions, head and body orientations, conversation groups of the subjects are annotated manually at 1/3 Hz. This dataset contains wearable sensor data captured by the Sociometric badges.

-*Conflab* [34]: contains 15 minute recordings of 49 people interacting with one another, captured by 5 (non-overlapping) overhead cameras. Positions, head and body orientations of the subjects are annotated manually at 60Hz. Note that even though locations and orientations can be acquired automatically, some previous works have pointed out that the automatic methods produce erroneous results, especially in orientation estimation [50]. To avoid confounding sources of errors in these behavioral cues which are quite nuanced as we motivated in Section 1, we follow other works that have relied on provided ground truth data as inputs assuming that these will be provided upstream during application [44]. Conversation groups are annotated manually at 1Hz.

4.3 Evaluation metrics

We evaluate on both stages of our model: (1) pairwise affinity estimation and (2) group detection. For pairwise affinity prediction, the neural network is trained with binary ground truth, but the predicted affinities are continuous values between 0 and 1, which enables us to do further analysis using these affinity scores. Given the dynamic nature of the proposed model, we introduce an additional evaluation compared to the existing state-of-the-art methods. Existing methods (e.g., [43, 45]) omit assessing the learned affinities only and use them directly for F-formation detection via DS clustering. We argue that there may be nuances in the learned pairwise continuous valued affinities that may anticipate changes in the group membership that may not be apparent from the hard cluster assignment.

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

For the second stage of evaluating group detection, we used the standard evaluation metric used in prior work [11, 23, 43, 47, 51] which involves considering an entire group in the ground truth as a single sample. A detected group k is considered to be correctly estimated if $\lceil Thr^* \|g_k\| \rceil$ of the members are correctly estimated, where $\|g_k\|$ indicates the cardinality or the size of the ground truth group, and $\lceil x \rceil$ rounds x to the next largest integer. The threshold $Thr \in [0, 1]$ tunes the tolerance of the evaluation to the number of mis-attributed members in a group. It is commonly set to $Thr = \frac{2}{3}$ or $Thr = 1$, representing greater than majority overlap at 67% and complete overlap with the ground truth membership, respectively. A True positive (TP) is therefore any correctly detected group; false negative (FN) is a missed group; and a false positive (FP) is an estimated group that does not exist in the ground truth. The metrics for group detection performance is then computed using F1 measure over the entire image scene which could contain multiple groups.

4.4 Implementation details

In the case of Conflab, the features were extracted to align with the ground truth at 1Hz by averaging all 60 samples before the label. While it is desired to use a higher frequency signal, this preprocessing step should already allow capturing of social dynamics that exist on a second level, such as synchrony and convergence patterns [24, 31], the associated postural sways [14], and some turn-taking dynamics (e.g., turn transitions). [49].

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

manually and the annotation method are described in each dataset respectively.

Similar to the experimental set up in DANTE, due to the small dataset size, all results are obtained by averaging the test splits using 5-fold cross validation. The validation split is selected such that it separates the training set as much as possible from the test data in time. The test data of a fold is used for results whereas hyperparameters are selected based on validation data. The hyperparameters are hidden representation dimension of the LSTM and sequence length of the input sequence. For the Conflab dataset, experimental results are obtained for all cameras (camera 2, 4, 6, and 8).

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

5 RESULTS AND DISCUSSION

5.1 Overview

To ensure a fairer comparison with existing methods, we use the same position and head and/or body orientation based feature set of the individuals in the scene for all methods. Table 1 shows an overview of the results on baseline methods on the Cocktail Party, SALSA, and Conflab datasets. As we expected, the proposed method outperforms the baseline methods on the Conflab dataset because of the finer temporal granularity in the behavioral cues and group labels. As opposed to DANTE which excels in both Cocktail Party and SALSA dataset, the proposed method may not have leveraged the temporal context when the social dynamics is undersampled.

Table 1: F1 scores comparison $Thr = 1$ across different methods. Standard deviation over test samples are shown in parenthesis. *: results reported by [43]; †: results reported by [45]

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

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

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

With increased performance in affinity estimations (i.e., the model output), the performance in Dominant Set clustering for group extraction also improves. As we argue that the affinity estimations are critical not only because they are inputs to DS clustering, but also contain valuable information on how pairwise relationships change continuously over time, we include a more detailed analysis of the affinity values and their relationship with the DS clustering step in the next sections.

5.2 Analysis of affinity values

To uncover where the difference in group detection results in Table 1 originates from, this section includes an analysis of where the proposed model differs from DANTE in terms of the predicted affinity value for test sets of the Cocktail Party and the Conflab dataset. Table 2 shows a comparison of the predicted affinity value results from DANTE and the proposed method using the AUC metric. With its data augmentation strategy and benefiting from the pairwise output setup, the frame-based DANTE strategy works better for the sparsely sampled Cocktail Party dataset. For the Conflab dataset with the higher sampling frequency, the proposed approach takes into consideration the temporal continuity of labelled cues and affinity values with the data-efficient person-wise training and output, resulting in improved AUC scores that led to improved conversation group detection F1 scores. We posit that the temporal granularity could be too coarse in datasets such as the Cocktail Party for the proposed sequential model to be effective.

Table 2: AUC results of DANTE and the proposed method for the Cocktail Party and Conflab dataset.

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

Figure 4 shows a qualitative example of how the affinity values from the proposed model change temporally as a new conversation group (Subject 2 and 3) forms. We focus on the right side of the interaction floor in the sequence of the scenes shown. The groups provided by the ground truth, predictions from the proposed method, and DANTE are illustrated in the second column. The affinity scores between Subject 1 and 2, and Subjects 2 and 3 from the proposed method are visualized in the third column. The color and value correspondence is shown in the legend. The pairwise affinity scores between Subject 1 and 2 decrease over time, whereas the score between Subject 2 and 3 increase over time.

5.3 Affinity scores in Dominant Set clustering

As the predicted affinity scores are continuous values, they are not perfect to directly extract group memberships. The pairwise values

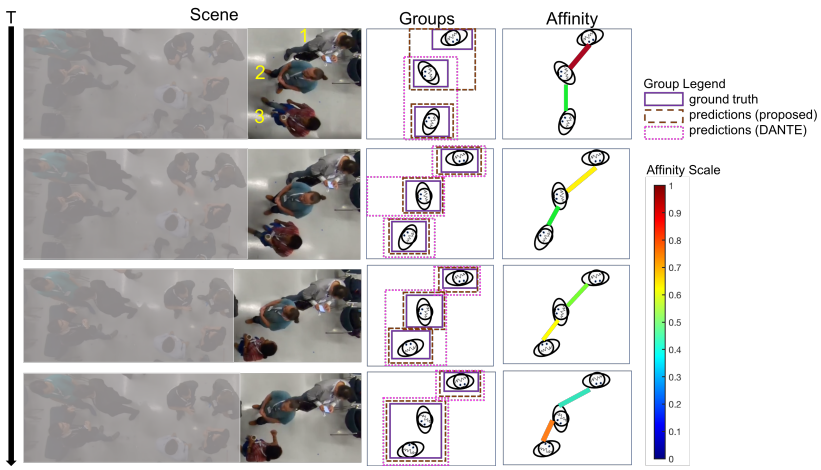

Figure 4: Qualitative example of how affinity values change temporally, in relation to a newly formed conversation group.

might differ and this results in discrepancy [21, 43]. Whether to symmetrize and how to symmetrize the predicted pairwise affinity scores is a design choice not thoroughly assessed previously. Options include using the asymmetric raw predicted affinity values, taking the minimum, average, or maximum of the pairwise affinity values. The (a)symmetry could be illustrative of the individual's intention in interacting with the other person, and affect the group clustering performance as this factor may have also affected how the annotators perceived group memberships.

In Table 3, we show the sensitivity F1 scores of symmetrizing the affinity matrix using different strategies for the Conflab dataset. The results show that averaging the pairwise affinities leads to improved F1 score in group detection at $Thr = 1$. This implies that while the asymmetric values are interesting in differentiating people's likelihood in interacting with each other, a symmetric and averaged representation is better aligned with the binary group membership.

Table 3: F1 scores comparison ($Thr = 1$) for the Conflab datset between different strategies of processing the affinity scores as inputs to DS clustering. Standard deviation over test samples are shwon in parenthesis.

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

5.4 Performance with respect to scene dynamics

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

Table 4: F1 scores comparison (Thr=1) for the Conflab dataset with respect to scene dynamics quantified by D. Standard deviation over test samples are shown in parenthesis.

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

We characterize the scenes in the Conflab dataset using this measure of scene dynamics. In Table 4, we showcase the performance difference (indicated by Delta) of group detection performance using F1 scores at $Thr = 1$ between DANTE and the proposed method. Delta is calculated by the F1 scores obtained from the proposed method substracted by that of DANTE. We show the results at different level of scene dynamics denoted as D, where D is the sum of all instances of group formations, breaking, and reformation. We observe a slight upward trend of the proposed method's improved performance (i.e., Delta) as scenes become more dynamic (with the exception of D=3). When $D \geq 5$ (corresponding to high scene dynamics, whereas $D = 0$ for most of the time where groups are stable), the advantage of the proposed method over DANTE on average is at 0.1. This further shows that the temporal context before a group event takes place may be beneficial in estimating conversation groups.

5.5 Conversation group forecasting

Using the temporal context of pairwise affinity scores, we further introduce a conversation group forecasting framework. Given a sequence of edge weights a_t^{edge} where $t = 0, 1, ..., T$, which is the averaged value between each pair of individuals i and j (i.e., $\frac{a_{i,j}+a_{j,i}}{2}$),

Table 5: Performance of forecasting conversation groups at different future time steps for the Conflab dataset (cam 6). Note that the column $t = T$ indicates the detection results. Uncertainty quantified by standard deviation across all scenes are shown in parenthesis.

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

we predict $a^{\text{edge}}_{t+1}, a^{\text{edge}}_{t+2}, ..., a^{\text{edge}}_{t+Z}$ where Z is the time forecast horizon. For each a^{edge}_{t} sequence, we fit a Gaussian Process Regressor (GPR) to provide uncertainty measure over the predictions.

GPR assumes a kernel that determines the covariance over target functions and uses the observations to obtain a likelihood function. A new posterior distribution can be computed based on Bayes' theorem. The choice of kernel characterized by a covariance function that measures the similarity between data points is an essential component in GPR. For the purpose of this study, this covariance function is chosen to be the popular Radial Basis Function (RBF). For more technical background on GPR, please refer to [35].

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

From the validation sets of the data, we found that a sequence length of 10 was the optimal hyperparameter for affinity prediction and hence, we set $T = 10$ to acquire corresponding observed samples of affinity scores to fit the GPR for this forecasting task. The forecast horizon Z represents the time steps beyond T (measured in seconds for the Conflab dataset). Note that the fitted GPR could be sampled continuously; we chose a set of discrete time steps beyond T for the scope of this paper. Table 5 shows the forecasting results of predicting the conversation groups in Conflab (cam 6) using the aforementioned approach. We report the averaged F1 scores from evaluating the N affinity matrix instances of each scene (i.e. aggregated from using the N samples of affinities drawn from each edge) for all scenes. The results show that there is a decreasing trend in the group detection performance as the forecast horizon extends, while the uncertainty in the group prediction in future scenes increases.

6 CONCLUSION AND FUTURE WORKS

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

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

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

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

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.

REFERENCES

[1] 2022. SPCL Midge badge. https://github.com/Jerzeek/spcl_midge_hardware

[2] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. 2016. SALSA: A Novel Dataset for Multimodal Group Behavior Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (Aug. 2016), 1707–1720. https://doi.org/10.1109/tpami.2015.2496269

[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*. 5–14.

[4] Lucy Baehren. 2022. Saying "goodbye" to the conundrum of leave-taking: a cross-disciplinary review. *Humanities and Social Sciences Communications* 9, 1 (2022), 1–13.

[5] Dan Bohus, Sean Andrist, and Eric Horvitz. 2018. A study in scene shaping: Adjusting F-formations in the wild.

[6] Dan Bohus and Eric Horvitz. 2009. Dialog in the open world. In *Proceedings of the 2009 international conference on Multimodal interfaces - ICMI-MLMI '09*. ACM Press. https://doi.org/10.1145/1647314.1647323

[7] Dan Bohus, Chit W Saw, and Eric Horvitz. 2014. Directions robot: in-the-wild experiences and lessons learned. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. Citeseer, 637–644.

[8] Samuel Rota Bulò and Marcello Pelillo. 2017. Dominant-set clustering: A review. *European Journal of Operational Research* 262, 1 (2017), 1–13.

[9] T Matthew Ciolek and Adam Kendon. 1980. Environment and the spatial arrangement of conversational encounters. *Sociological Inquiry* 50, 3-4 (1980), 237–271.

[10] Joe Connolly, Nathan Tsoi, and Marynel Vázquez. 2021. Perceptions of Conversational Group Membership based on Robots' Spatial Positioning: Effects of Embodiment. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. ACM. https://doi.org/10.1145/3434074.3447195

[11] 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 *Procedings of the British Machine Vision Conference 2011*. British Machine Vision Association. https://doi.org/10.5244/c.25.23

[12] Marco Cristani, Ramachandra Raghavendra, Alessio Del Bue, and Vittorio Murino. 2013. Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing* 100 (2013), 86–97.

[13] 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.

[14] Tetsuro Funato, Shinya Aoi, Nozomi Tomita, and Kazuo Tsuchiya. 2016. Smooth enlargement of human standing sway by instability due to weak reaction floor and noise. *Royal Society Open Science* 3, 1 (2016), 150570.

[15] Tian Gan, Yongkang Wong, Daqing Zhang, and Mohan S Kankanhalli. 2013. Temporal encoded F-formation system for social interaction detection. In *Proceedings of the 21st ACM international conference on Multimedia*. 937–946.

[16] 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.

[17] Ilka H Gleibs, Neil Wilson, Geetha Reddy, and Caroline Catmur. 2016. Group dynamics in automatic imitation. *PloS one* 11, 9 (2016), e0162880.

[18] Erving Goffman. 1955. On face-work: An analysis of ritual elements in social interaction. *Psychiatry* 18, 3 (1955), 213–231.

[19] Erving Goffman. 1967. Interaction ritual: Essays on face-to-face interaction. (1967).

[20] Edmund T Hall and Edward Twitchell Hall. 1966. *The hidden dimension*. Vol. 609. Anchor.

[21] Hooman Hedayati, Daniel Szafir, and Sean Andrist. 2019. Recognizing F-Formations in the Open World. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. https://doi.org/10.1109/hri.2019.8673233

[22] Michael J Hove and Jane L Risen. 2009. It's all in the timing: Interpersonal synchrony increases affiliation. *Social cognition* 27, 6 (2009), 949–960.

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

[24] Öykö Kapcak, Jose Vargas-Quiros, and Hayley Hung. 2019. Estimating Romantic, Social, and Sexual Attraction by Quantifying Bodily Coordination using Wearable Sensors. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 154–160.

[25] Kleomenis Katevas, Katrin Hänsel, Richard Clegg, Ilias Leontiadis, Hamed Haddadi, and Laurissa Tokarchuk. 2019. Finding Dory in the Crowd. In *Proceedings of the 1st Workshop on Machine Learning on Edge in Sensor Systems - SenSys-ML 2019*. ACM Press. https://doi.org/10.1145/3362743.3362959

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

[27] Annica Kristoffersson, Silvia Coradeschi, Kerstin Severinson Eklundh, and Amy Loutfi. 2013. Towards measuring quality of interaction in mobile robotic telepresence using sociometric badges. *Paladyn, Journal of Behavioral Robotics* 4, 1 (2013), 34–48.

[28] Oren Lederman, Akshay Mohan, Dan Calacci, and Alex Sandy Pentland. 2018. Rhythm: A Unified Measurement Platform for Human Organizations. *IEEE MultiMedia* 25, 1 (Jan. 2018), 26–38. https://doi.org/10.1109/mmul.2018.112135958

[29] Nicolai Marquardt, Robert Diaz-Marino, Sebastian Boring, and Saul Greenberg. 2011. The proximity toolkit. In *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*. ACM Press. https://doi.org/10.1145/2047196.2047238

[30] 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 (March 2018), 1–30. https://doi.org/10.1145/3191757

[31] Jose David Vargas Quiros, Oyku Kapcak, Hayley Hung, and Laura Cabrera-Quiros. 2021. Individual and joint body movement assessed by wearable sensing as a predictor of attraction in speed dates. *IEEE Transactions on Affective Computing* (2021).

[32] Navin Raj Prabhu, Chirag Raman, and Hayley Hung. 2020. Defining and quantifying conversation quality in spontaneous interactions. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 196–205.

[33] 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.

[34] Chirag Raman, Jose Vargas-Quiros, Stephanie Tan, Ekin Gedik, Ashraful Islam, and Hayley Hung. 2022. ConfLab: A Rich Multimodal Multisensor Dataset of Free-Standing Social Interactions In-the-Wild. https://doi.org/10.48550/ARXIV.2205.05177

[35] Carl Edward Rasmussen. 2003. Gaussian processes in machine learning. In *Summer school on machine learning*. Springer, 63–71.

[36] 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 (ICCV)*.

[37] 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.

[38] Jorge Rios-Martinez, Anne Spalanzani, and Christian Laugier. 2015. From proxemics theory to socially-aware navigation: A survey. *International Journal of Social Robotics* 7, 2 (2015), 137–153.

[39] Viktor Schmuck and Oya Celiktutan. 2021. GROWL: Group Detection With Link Prediction. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 1–8.

[40] Francesco Setti, Oswald Lanz, Roberta Ferrario, Vittorio Murino, and Marco Cristani. 2013. Multi-scale f-formation discovery for group detection. In *2013 IEEE International Conference on Image Processing*. IEEE. https://doi.org/10.1109/icip.2013.6738732

[41] 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.

[42] Mariëlle Stel, Jim Blascovich, Cade McCall, Jessanne Mastop, Rick B Van Baaren, and Roos Vonk. 2010. Mimicking disliked others: Effects of a priori liking on the mimicry-liking link. *European Journal of Social Psychology* 40, 5 (2010), 867–880.

[43] 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. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 20 (may 2020), 23 pages. https://doi.org/10.1145/3392824

[44] Stephanie Tan, David M. J. 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 (March 2021), 1–22. https://doi.org/10.1145/3448122

[45] 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*. ACM. https://doi.org/10.1145/3462244.3479963

[46] Bruce W. Tuckman. 1965. Developmental sequence in small groups. *Psychological Bulletin* 63, 6 (1965), 384–399. https://doi.org/10.1037/h0022100

[47] 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.

[48] Marynel Vázquez, Elizabeth J Carter, Braden McDorman, Jodi Forlizzi, Aaron Steinfeld, and Scott E Hudson. 2017. Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 42–52.

[49] Margaret Wilson and Thomas P Wilson. 2005. An oscillator model of the timing of turn-taking. *Psychonomic bulletin & review* 12, 6 (2005), 957–968.

[50] 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.

[51] Lu Zhang and Hayley Hung. 2016. Beyond f-formations: Determining social involvement in free standing conversing groups from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1086–1095.

[52] Lu Zhang and Hayley Hung. 2018. On social involvement in mingling scenarios: Detecting associates of f-formations in still images. *IEEE Transactions on Affective Computing* 12, 1 (2018), 165–176.