Improving Social Awareness Through DANTE: A Deep Affinity Network for Clustering Conversational Interactants

MASON SWOFFORD*, Stanford Vision Lab
JOHN PERUZZI*, Stanford Vision Lab
NATHAN TSOI, Yale Interactive Machines Group
SYDNEY THOMPSON, Yale Interactive Machines Group
ROBERTO MARTIN-MARTIN, Stanford Vision Lab
SILVIO SAVARESE, Stanford Vision Lab
MARYNEL VÁZQUEZ, Yale Interactive Machines Group

We propose a data-driven approach to detect conversational groups by identifying spatial arrangements typical of these focused social encounters. Our approach uses a novel Deep Affinity Network (DANTE) to predict the likelihood that two individuals in a scene are part of the same conversational group, considering their social context. The predicted pair-wise affinities are then used in a graph clustering framework to identify both small (e.g., dyads) and large groups. The results from our evaluation on multiple, established benchmarks suggest that the combination of powerful deep learning methods with classical clustering techniques can improve the detection of conversational groups in comparison to prior approaches. Finally, we demonstrate the practicality of our approach in a human-robot interaction scenario. Our efforts show that our work advances group detection not only in theory, but also in practice.

CCS Concepts: • **Computing methodologies → Spatial and physical reasoning**; • **Human-centered computing → Collaborative and social computing systems and tools**.

Additional Key Words and Phrases: group conversations, spatial analysis, proxemic interactions, f-formations

ACM Reference Format:
Mason Swofford, John Peruzzi, Nathan Tsoi, Sydney Thompson, Roberto Martin-martin, Silvio Savarese, and Marynel Vázquez. 2018. Improving Social Awareness Through DANTE: A Deep Affinity Network for Clustering Conversational Interactants. *J. ACM* 37, 4, Article 111 (August 2018), 22 pages. https://doi.org/10.1145/1122445.1122456

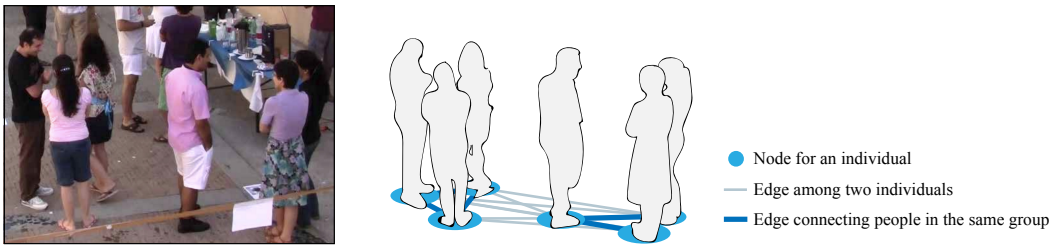
1 INTRODUCTION

Automatic detection of group conversations enables a rich set of intelligent, social computer interfaces. For example, group detection has traditionally enabled surveillance systems [17, 31, 57], socially-aware mobile systems [41], interactive displays [33], exhibits [20], and social playgrounds

*Both authors contributed equally to this research.

Authors' addresses: Mason Swofford, Stanford Vision Lab, 353 Serra Mall, Stanford, California, 94305, mswoff@stanford.edu; John Peruzzi, Stanford Vision Lab, jperuzzi@stanford.edu; Nathan Tsoi, Yale Interactive Machines Group, 51 Prospect St, New Haven, Connecticut, 06511, nathan.tsoi@yale.edu; Sydney Thompson, Yale Interactive Machines Group, sydney.thompson@yale.edu; Roberto Martin-martin, Stanford Vision Lab, roberto.martinmartin@stanford.edu; Silvio Savarese, Stanford Vision Lab, ssilvio@stanford.edu; Marynel Vázquez, Yale Interactive Machines Group, marynel.vazquez@yale.edu.

0004-5411/2018/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456



(a) Scene from the Coffee Break dataset (b) Interaction Graph

Fig. 1. Example problem setting. Given a social scene (a), the goal is to identify the individuals that are part of the same conversational group based on their position and orientation relative to one another (b). We approach this problem by combining graph clustering with deep learning. See the text for more details.

[35, 45]. In the context of robotics, group detection is also essential for situated spoken language interaction [9], non-verbal robot behavior generation [64], and socially-aware robot navigation in human environments [53]. However, detecting conversations in dynamic human environments is an intricate problem, requiring the perception of subtle aspects of social interactions.

In this work, we study the problem of visually recognizing situated group conversations by analyzing *proxemics* – people's use of physical space [29]. In particular, we study automatic recognition of spatial patterns of human behavior that naturally emerge during group conversations [25]. These patterns are known as Face Formations, or *F-Formations* in short, as denoted by A. Kendon [36]. They are the result of people needing to communicate in close proximity while sustaining a shared, focus of attention. Prototypical formations are often observed as face-to-face, side-by-side or circular spatial arrangements in open spaces. However, the specific type of arrangement that emerges during conversations ultimately depends on a number of social factors, including the number of interactants, their conversation topic, and environmental spatial constraints. Worth noting, F-Formations provide a conceptual framework in Human-Computer Interaction (HCI) for thinking about how the physical aspects of a setting influence interactions [4, 19, 42, 59].

Most prior work on visual F-Formation detection has focused on explicitly modeling properties of conversational group spatial arrangements [31, 56, 57]. For instance, people tend to keep a social distance from one another during conversations [29] and orient their bodies towards the center of their group [36]. But these approaches do not typically account for the malleability inherent in human spatial behavior. For example, people naturally adapt to crowded environments and modify their spatial formations by interacting closer if need be. Robustness to these complex scenarios is essential for reasoning about group conversations through spatial analysis in real applications.

Contrary to most prior work on F-Formation detection, we explore using the powerful approximation capabilities of Deep Learning (DL) for identifying conversations and their members. But how can one leverage these approximation capabilities effectively given the small-sized datasets that are available for F-Formation detection? Additionally, how can one deal with variable number of group interactants? To answer these questions, we revisit ideas from classical graph clustering solutions [31, 62, 67]. We view the F-Formation detection problem as finding sets of related nodes in an *interaction graph* (Figure 1). The nodes of the graph correspond to individuals in a scene with associated spatial features obtained through image processing. The graph edges connect two nearby people and have an associated affinity (weight) that encodes the likelihood that they are conversing. Under this framing, the key challenge for F-Formation detection is to compute appropriate affinities for identifying groups. While prior work used simple heuristics to compute edge weights [31, 62, 67], we propose to learn a function that predicts these weights.

Our main contributions are threefold:

(1) We propose a novel **D**eep **A**ffinity **N**e**T**work for clust**E**ring conversational interactants (DANTE). The network approximates the likelihood that two individuals are conversing given their spatial features and information about other nearby individuals, i.e., their social context. The network can deal with contexts of varying sizes by leveraging recent ideas to input sets to DL [50].
(2) We conduct an extensive evaluation of DANTE in established benchmarks. The evaluation shows that our proposed approach advances conversational group detection in comparison to state-of-the-art methods. Furthermore, our results show that DANTE is extensible. It can easily reason about different relevant spatial features for group detection.
(3) Finally, we demonstrate the applicability of DANTE on a social robotic system. In this context, DANTE enables a robot to identify the members of its group conversations as well as nearby bystanders, who offer opportunities for new interactions. We open-source our code to facilitate future replication efforts and enable the use of DANTE to create other socially-aware interfaces.

2 RELATED WORK

A variety of approaches have been proposed to computationally detect situated group interactions through proxemics analysis within HCI [10, 15, 30, 41], computer vision [3, 22, 27, 47], social signal processing [13, 46], and even natural language processing [66]. Due to limited space, this section focuses on describing close related work on F-formation detection based on visual spatial features. For a broader review, we encourage interested readers to refer to [1] and [63] (Section 3.2).

Visual conversational group detection. Our work focuses on the analysis of visual human spatial behavior because: (1) visual sensing with cameras is cheap and does not require instrumentation of users, thus enabling group detection in unconstrained settings [8]; (2) relevant visual features are readily available through open-source or commercial software [11, 44], (3) prior work has shown that these visual features are effective to enable proxemics interactions [40]. It is worth noting that early research on conversational group detection within the computer vision community was motivated by surveillance applications in public human environments [12, 14, 24, 67]. These approaches identified two key features for spatial analysis: human *position* and *orientation* information. We also use these features in our work.

Most prior approaches for visually detecting F-Formations are based on mathematical models of sustained spatial arrangements [16, 23, 56, 57, 65]. These model-based approaches tend to formalize the *transactional segments* of individuals, which are the space that extends forward from their lower body and that includes whatever they are currently engaged with. Then, these methods find the intersection of transactional spaces in a scene, or *o-spaces* of the F-Formations [36]. For example, [16, 56, 65] use voting schemes to find o-spaces, while Setti et al. [57] use an iterative graph-cuts approach. We consider the latter work [57] in our evaluation as it provides state-of-the-art results on group detection using spatial features from a single image.

Group detection as graph clustering. An important category of group detection methods rely on graph clustering [31, 62, 67], including ours. In this setting, individuals correspond to nodes in a weighted, interaction graph (Fig. 2b). The goal is to partition the graph into groups of nodes that represent human interactions. Note that soft group assignments are also possible [12, 65], but we focus on the hard assignment problem in this work because standard evaluation benchmarks provide hard group labels.

Similar to [31, 62], we use the Dominant Sets clustering algorithm to detect F-Formations in social scenes. Different to these prior efforts, though, we do not use hand-crafted heuristics [31] nor a model of human attention [62] to assign weights to graph edges and perform clustering. Instead, we propose to learn in a data-driven fashion a non-linear function that predicts the weights. One of

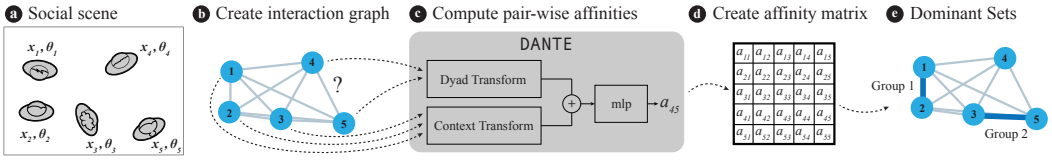


Fig. 2. Group detection approach. Our method receives as input spatial features (e.g., position $\mathbf{x}$ and orientation θ) for the social agents in a scene (a). This information is used to create an interaction graph (b) and to compute pair-wise affinities with DANTE (c). The affinities are assembled into an affinity matrix (d) to cluster nodes (e). Multi-layer perceptron is abbreviated as mlp in (c). See the text for more details.

our insights in this prediction problem is to consider the spatial information of individuals nearby the corresponding dyad of interest for a graph edge, i.e., to consider the dyad's *social context*.

Deep learning (DL) applied to group detection. Two recent methods have attempted to use DL for conversational group detection by reasoning about spatial information. First, [58] leveraged DL for estimating the location of o-spaces in a scene. Then, this prior method used a greedy geometrical approach to assign interactants to group conversations. We omit evaluating our approach against [58] since their results were far below the state-of-the-art [57]. Second, Sanghvi, Yonetani and Kitani [54] proposed to use DL in the context of learning communication policies. As part of their policy network, the authors introduce a communication gating module that automatically infers group membership. We consider this approach as a baseline in our evaluation given that their results are comparable to [57].

Group detection evaluation. We evaluate our approach on standard datasets for group detection within the computer vision and social signal processing community [2, 5, 16, 68] – to the best of our knowledge, there are no other datasets available for the visual group detection task in HCI. The datasets provide ground truth group annotations, as well as position and head orientation for the individuals in the scene. The latter features were gathered with automated computer vision techniques, thus providing realistic inputs for our experimental evaluation.[1]

Several prior efforts have demonstrated the value of F-formation detection in HCI. For example, F-formation detection has enabled spatial interfaces [4] and the generation of coordinated, non-verbal robot behavior in situated human-robot conversations [64]. Likewise, we demonstrate the applicability of our proposed approach in a real interactive system. This effort reinforces the value of F-Formation detection for building socially-aware interactive systems.

3 CONVERSATIONAL GROUP DETECTION WITH DANTE

Our group detection method is illustrated in Fig. 2. The input to our method is a set of N potential interactants in a scene $\mathcal{I} = \{i_0, \ldots, i_N\}$. These interactants are typically people but could also be other social agents relevant for spatial analysis, like robots [32, 38, 64]. Each $i_i = (id_i, \mathbf{f}_i)$ in $\mathcal{I}$ has a unique identifier id_i for the social agent and a feature vector $\mathbf{f}_i$ with corresponding spatial information. By default, we include in the vector $\mathbf{f}_i$ the 2D position of the individual in the planar layout of the environment, $\mathbf{x}_i = (x_i, y_i)$, and its orientation, θ_i, relative to a world coordinate frame (Fig. 2(a)). Our rationale for including these features stems from prior work both in social psychology [25, 36] and computer science [12, 16, 62] that have shown the importance of these features for modeling F-Formations. But more generally, $\mathbf{f}_i$ could also include other spatial information, like the

[1]Even though A. Kendon [36] defined transactional segments based on people's lower body orientation [36], we often use head orientation as a key feature for group detection. Our rationale for this decision is evaluating our approach using established datasets [16, 68]. As in [52], though, we believe that future work should consider both body and head orientation for interaction analysis. Our method could easily be extended to this end.

interactant's instantaneous velocity, as later discussed in our evaluation. The output of our method is another set, but in this case of K detected conversational groups, $\mathcal{G} = \{g_0, \ldots, g_K\}$. Each group g_k is composed of the identifiers of the interactants that belong to it, $g_k = \{id_i\}$. The conversational groups are mutually exclusive: a social agent can only belong to a single conversational group.

3.1 Approach

Our approach represents the scene as an interaction graph $G = (V, E, A)$ with a set of nodes or vertices V, edges E, and non-negative affinities or weights A. As shown in Fig. 2(b), each node corresponds to a social agent and contains its data $i_i = (id_i, \mathbf{f}_i)$. Pairs of nodes are connected by undirected edges in the graph. For each edge, its affinity score is meant to represent the likelihood that the two agents connected to the edge belong to the same group conversation.

The main insight of our work is learning the graph affinities such that we can effectively partition the graph to determine the groups $\mathcal{G}$. In particular, we propose DANTE, which stands for Deep Affinity NeTwork for clustEring interactants, to predict affinity scores (Fig. 2(c)). The benefits of DANTE include reducing the reliance of group detection on heuristics in comparison to prior work, e.g., [31, 57, 62, 65, 67]. Section 3.2 further explains DANTE in detail.

Our proposed F-Formation detection approach finally builds an affinity (or similarity) matrix to cluster nodes with the Dominant Sets (DS) algorithm [31] (Fig. 2(d)-(e)). The DS algorithm iteratively finds clusters that describe compact structures, which are well suited to represent F-Formations. More details on this graph clustering procedure are provided in Sec. 3.3.

3.2 Affinity Scoring with DANTE

This section introduces DANTE, a neural network that predicts the weights for each of the edges in the social graph (Fig. 2c). DANTE is structured to reason about two types of information: local spatial information from the two vertices (individuals) connected to an edge of interest, and global spatial information from other nearby people, i.e., the social context of the dyad of interest. Predicting affinities based on these two types of information within a graph clustering framework is a novel contribution of our work. Because of it, our approach does not need additional ad-hoc steps [57, 65] to verify that the detected groups effectively conform with the notion of F-Formations [36]. DANTE's structure and its inherent data-driven nature make it easily extensible to applications with a variety of spatial features, as demonstrated in our evaluation.

Without loss of generality, assume for the following sections that DANTE is computing the affinity a_{ij} for the individuals i and j in the interaction graph G.

3.2.1 Dyad Transform. The Dyad Transform of DANTE is in charge of computing *local features* for i and j, as depicted in the top part of Fig. 3. The input to the Dyad Transform is a matrix with two rows, one for the feature $\mathbf{f}_i$ and one for $\mathbf{f}_j$. Each of these features are transformed independently by a multi-layer perceptron (mlp), resulting in a feature encoding for each social agent of dimensionality d_{dyad}. The mlp is composed of D_{dyad} dense layers followed by ReLU activations. The result is a matrix of features in $\mathbb{R}^{2\times d_{dyad}}$. This matrix is finally flattened into a dyad feature vector $\mathbf{v}_{dyad} \in \mathbb{R}^{1\times 2d_{dyad}}$.

3.2.2 Context Transform. DANTE's Context Transform computes a *global feature* representation for the social context of the dyad of interest, as illustrated in the bottom part of Fig. 3. Our design for this model component is inspired by prior work on inputting sets to neural networks, especially in the context of point cloud processing [50]. Our experimental results reinforce the notion that neural networks can encode spatial features in a scalable manner for social interaction analysis [28, 49].

At its core, the Context Transform uses a symmetric function to handle unordered and potentially variable number of inputs. In our case, these inputs correspond to the set $\mathbf{F}$ of individual, spatial

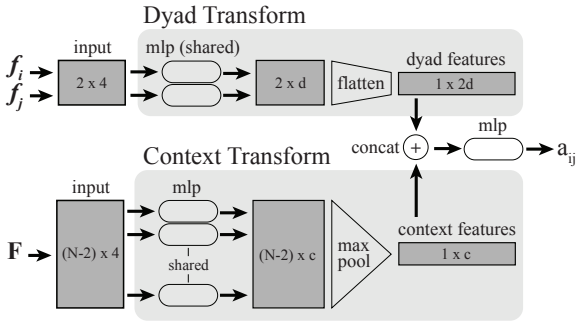


Fig. 3. DANTE components. The pairwise affinity a_{ij} of a pair of individuals i and j is computed from two types of features: the local *dyad features* and the global *context features*. As explained in Sec. 3.2.4, all spatial data that is input to DANTE is transformed to the canonical frame between i and j before computation. The abbreviations "concat" and "mlp" stand for concatenation and multi-layer perceptron, respectively.

feature vectors for the people in the scene other than the dyad of interest. More formally, assume again that DANTE is computing the affinity a_{ij}. Then, the input to the Context Transform is a feature set $\mathbf{F} = \{\mathbf{f}_k \mid 1 \leq k \leq N \text{ and } k \notin \{i, j\}\}$ with each vector $\mathbf{f}_k$ encoding the agent's k spatial information (e.g., position and orientation) as explained in previous sections. The set $\mathbf{F}$ is implemented as a matrix with one feature per row.

Similarly to the Dyad Transform, the Context Transform first applies independently a multi-layer perceptron to each of the rows of its input matrix $\mathbf{F}$ (bottom part of Fig. 3). The mlp is composed of $D_{context}$ dense layers followed by ReLU activations, resulting in a matrix of features in $\mathbb{R}^{(N-2)\times d_{context}}$. The latter matrix is finally transformed by max pooling along its rows. The output is a context feature vector $\mathbf{v}_{context} \in \mathbb{R}^{1\times d_{context}}$. Note that max pooling is the key symmetric operation that makes the Context Transform invariant to input permutations.

3.2.3 Combining Dyad and Context Features. Finally, DANTE combines the dyad information $\mathbf{v}_{dyad}$ and context information $\mathbf{v}_{context}$ to compute the affinity score a_{ij}. To this end, DANTE first concatenates the two feature vectors column-wise, resulting in a new vector in $\mathbb{R}^{1\times(2d_{dyad}+d_{context})}$. Then, an mlp is used to transform the combined features into a joint representation. In this case, the mlp is composed of D_{comb} dense layers, each followed by ReLU activations. Finally, one more dense layer projects down the resulting features into a scalar value. This last layer uses a sigmoid activation function to constraint the output to the $[0, 1]$ range.

3.2.4 Other Implementation Details. The spatial features $\mathbf{f}_i$ are originally obtained in a world reference frame W. However, we transform them before inputting them to DANTE to a canonical frame of reference defined with respect to the pair of individuals whose affinity a_{ij} is being computed. This canonical frame O_{ij} is illustrated in Fig. 4. The frame O_{ij} is located at the middle point between the social agents $(\mathbf{x}_i + \mathbf{x}_j)/2$ in the global frame W. For setting orientation of O_{ij}, we align its x axis with a vector from the position of i to the position of j. This transformation facilitates learning and generalization.

By default, we use 4-dimensional representation for the spatial features $\mathbf{f}_i$ relative to O_{ij} in our implementation (Fig. 3). The four dimensions correspond to the 2D position $(\mathbf{x}_i)$ and the orientation θ_i of the social agent, with the angle encoded through $\sin(\theta_i)$ and $\cos(\theta_i)$. Using sine and cosine helps avoid issues with θ_i wrapping around 360°. Another benefit is that the projection forward in the direction that an agent is looking is the result of a simple multiplication of the sine and cosine of

the orientation by some positive value. This projection has been employed by several other group detection algorithms [16, 56, 57], suggesting that it can facilitate reasoning about F-Formations. We theorize that the proposed representation can be used by DANTE to easily learn to process spatial data in a useful manner. Furthermore, we show in Sec. 5 how DANTE can be adapted to alternative spatial feature representations when orientation information is not readily available.

3.3 Dominant Sets Grouping

Once the affinities for each pair of individuals in the social interaction graph are computed, our approach proceeds to group people using the Dominant Sets (DS) algorithm by Hung and Kröse [31]. Dominant sets (clusters) in the algorithm are a generalization of maximal cliques to edge-weighted graphs with no self-loops [48]. Our social interaction graph G is one such graph.

For a detailed explanation of the Dominant Sets algorithm, we refer interested readers to Sections 3.2 and 6 of [31]. In short and for completeness, the DS algorithm iteratively finds clusters that satisfy the following property: the mutual affinity, formally defined in [31], between all of the cluster members is higher than the affinity between its members and those outside of it. Once a cluster is found, new clusters that satisfy the same property are then searched for. The stopping criterion for finding clusters in the DS algorithm is reached when (a) a new cluster either does not satisfy the property of high relative mutual group affinity or (b) when the mutual affinity of a group is below a certain threshold, which is determined through cross validation. In general, the DS algorithm results in compact clusters that are well suited to represent F-Formations of any size.

Although DS can be applied to social interaction graphs with asymmetric affinities, symmetric affinities have been reported to yield superior results [31, 62]. Thus, we assume symmetric affinities in our work by setting edge weights to the average of the predicted a_{ij} and a_{ji}, for $1 \leq i, j \leq N$ and $i \neq j$. Note that we could have made DANTE directly output symmetric affinities by making the Dyad Transform permutation invariant, like the Context Transform. However, we found in our initial tests that this design led to reduced performance in comparison to averaging a_{ij} and a_{ji}.

4 EVALUATION

We conduct systematic evaluations of our proposed group detection approach in established benchmarks. In this Section in particular, we first describe experiments with datasets that were created specifically for the evaluation of conversational group detection algorithms. These datasets provide position and orientation information for the individuals in a scene, as needed for our default spatial feature representation (Sec. 3.2.4). In Section 5, we further evaluate DANTE in a different setup to test its generalization capabilities to the more general problem of group detection. This involves the detection of F-Formations, but potentially also other types of group formations. Worth noting, the Appendix provides an additional experiment with synthetic data, which is not discussed outside Sec. A.2 because results were inconclusive.

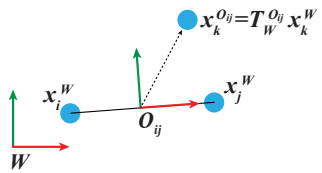


Fig. 4. When computing a given affinity a_{ij}, we transform the features input to DANTE from the world frame W to a local frame O_{ij} relative to the two potential interactants of interest (i, j).

4.1 Datasets

We consider three publicly available datasets of social interactions in our evaluation, presented below in order of annotation quality:

- *Cocktail Party Dataset* [68]. Contains about 30 min. of video recordings of a cocktail party in a lab environment. The video shows 6 people conversing with one another and consuming drinks and appetizers. The party was recorded using four synchronized cameras installed in the corners of the room. SubjectsâĂŹ positions were logged using a particle filter-based body tracker with head pose estimation [39]. Conversational groups were annotated at 5 sec. intervals, resulting in 320 frames with ground truth group annotations.
- *SALSA Dataset* [2]. 18 participants were recorded using multiple cameras and sociometric badges and then annotated at 3 second intervals over the course of 60 minutes, giving 1,200 total frames. The dataset consists of a *poster presentation* session and a *cocktail party*. Despite the differences in the structure of F-Formations that appear in these two settings, we treat SALSA as a single dataset to test generalization to different group formations.
- *Coffee Break Dataset* [16]. Images were collected using a single camera outdoors. People engaged in small group conversations during coffee breaks. The number of people per frame varied from 6 to 14. People tracking is rough, with orientations only taking values of 0, 1.57, 3.14, and 4.71 radians. Compared to Cocktail Party and SALSA, the spatial features provided by Coffee Break are far noisier. A total of 119 frames have ground truth group annotations.

4.1.1 Data Augmentation. Due to the small size of the datasets, we augment them during training. We flip position and orientation data over the horizontal and vertical axes of the world coordinate frame W, giving four times as many training examples. Since groups are define based on person ID's, they do not need to be adjusted for the augmented data.

4.2 Evaluation Metrics

We consider standard evaluation metrics for conversational group detection [16, 54, 56, 57, 62]. A given group k is said to be correctly estimated if $\lceil T * |g_k| \rceil$ of their members are correctly estimated and if no more than $1 - \lceil T * |g_k| \rceil$ false subjects are identified, where $|g_k|$ is the cardinality of the labeled group g_k and T is a defined tolerance threshold. Common values of T are $2/3$ and 1 [16, 54, 56, 57, 62, 65]. We center our attention on evaluating methods based on $T = 1$, i.e., on perfect group detection, since it is more challenging than $T = 2/3$.

Let TP (true positive) to be a correctly detected group, FN (false negative) be a non-detected group, and FP (false positive) be a group that was detected but did not exist. Then, we measure our accuracy with three metrics: precision, recall, and F_1 score. Precision is defined as $Prec = \frac{TP}{TP+FP}$, recall is $Rec = \frac{TP}{TP+FN}$, and F_1 score is $F_1 = 2 \cdot \frac{Prec \cdot Rec}{Prec+Rec}$.

4.3 Baselines

We focus on comparing the proposed conversational group detection approach against two state-of-the-art methods. First, we compare our proposed approach against the graph-cuts method by Setti et al. [57] (GCFF), given that it outperforms prior model-based group detection approaches, e.g., [16, 56]. Second, we compare results against the game theoretic approach of Vascon et al. [62] (GTCG) because this method relies on graph clustering, like our approach. Also, [62] tends to give better performance than the alternative graph-based approach by Hung and Kröse [31].

Although we did not re-run the group communication approach by Sanghvi, Yonetani, and Kitani [54] (GComm), we report results from their publication as a reference. This approach is of interest as well because it uses deep learning for conversational group detection.

4.4 Experimental Setup

Due to the small size of the datasets, we use 5-fold cross validation to measure performance and study variability in the results. Each fold is taken as a continuous section of data due to the inherent auto-correlation of time-series spatial features from the datasets' videos. We select data for validation from the training set such that it separates as much as possible the data that is actually used for training from the one that is used for testing. The test data of a given fold is only used for computing final results after hyper-parameters are chosen based on the validation set.

In order to fairly compare our results against previous work, we fine-tuned the state-of-the-art baselines [57, 62] using the corresponding training and validation data for each fold. The average results for the graph-cuts approach of Setti et al. were slightly improved in comparison to [57]. Note that [62] does not present results for $T = 1$ F1.

To train DANTE, we use the log loss between each predicted affinity and the true $\{0, 1\}$ affinity, corresponding to whether or not two people are part of the same conversational group. We optimize the network through gradient descent with the Adam optimizer [37], a learning rate of 0.0001 and a batch size of 64 samples. We search for the best hyper-parameters using the validation data, including the number of layers in DANTE's multi-layer perceptrons (D_{dyad}, $D_{context}$, D_{comb}) as well as the size of these layers.

4.5 Quantitative Results - Comparison Against Baselines

Table 1 shows quantitative results for the Cocktail Party, SALSA, and Coffee Break datasets. The results correspond to $T = 1$ F1 scores, averaged across all five folds. See the Appendix for expanded results per fold.

On average, our proposed approach outperforms the F1 scores of the baselines in the Cocktail Party dataset (see DANTE row in Table 1). The average improvement is 9% over GCFF, 13% over GComm, and 44% over GTCG. We conducted pairwise t-tests comparing per fold results of our approach against GTCG, and found that DANTE led to significantly higher F1 scores ($t(6.88) = 5.08$, $p = 0.002$). The difference between DANTE and GCFF was not statistically different ($p = 0.28$) but DANTE clearly outperformed GCFF in the first and second folds (see Table 3 in the Appendix). These results suggest that our approach improves the state of the art in conversational group detection when the input spatial data has reasonable quality.

Our proposed approach significantly outperforms GTCG ($t(5.36) = 3.93$, $p = 0.01$) and GCFF ($t(6.7) = 2.59$, $p = 0.04$) in the SALSA dataset; the comparison was not possible against GComm because [54] does not evaluate on the SALSA dataset. Across all folds, our method results in higher F-1 scores than the baselines (see Table 4 in the Appendix). This finding suggests that our data-driven approach can handle the different types of group formations observed in SALSA.

The benefits of our approach are less noticeable in the Coffee Break dataset, where our method performs as well as the best baselines (GCFF & GComm). Although DANTE leads on all folds in

Table 1. Results on various conversational group detection benchmarks. The scores are $T = 1$ F1 values averaged across five folds. See the Appendix for a break-down of the results by each fold.

Method	Cocktail Party	Coffee Break	SALSA
GComm [54]	0.60	0.63	-
GTCG [62]	0.29	0.48	0.44
GCFF [57]	0.64	0.63	0.41
DANTE	**0.73**	0.64	**0.65**
DANTE-NoContext	0.64	**0.66**	0.57

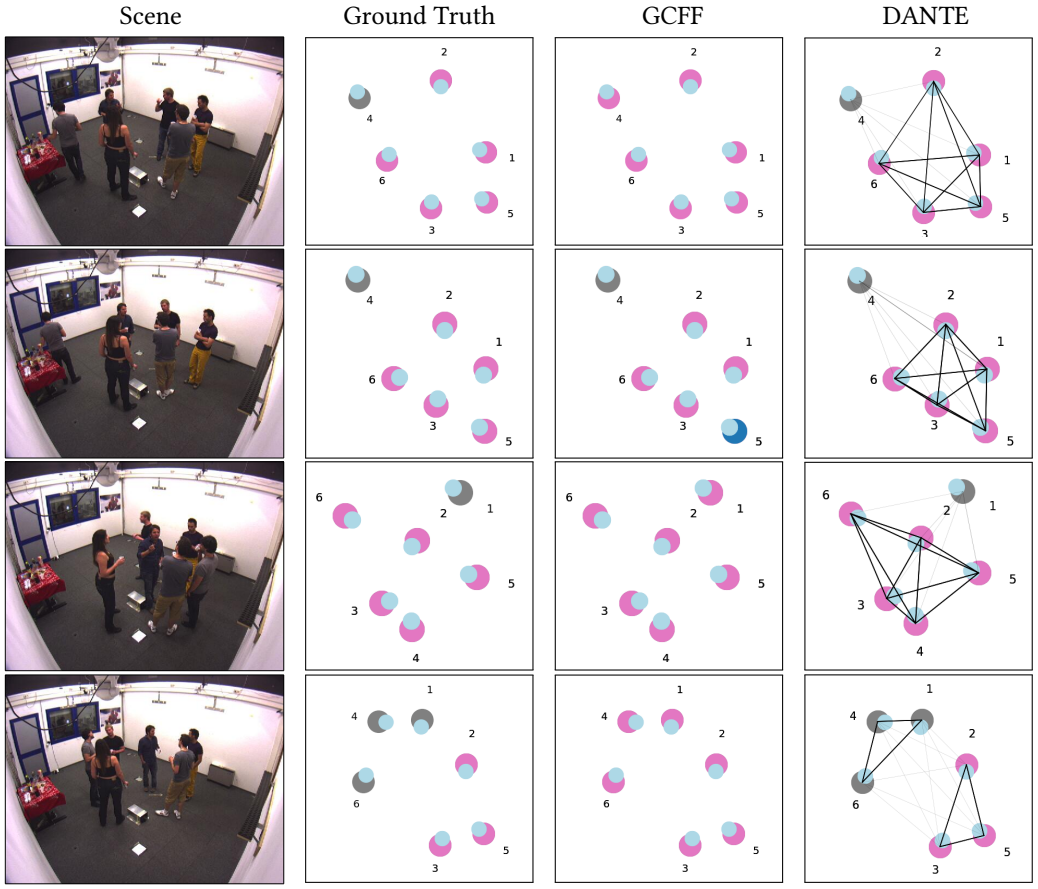
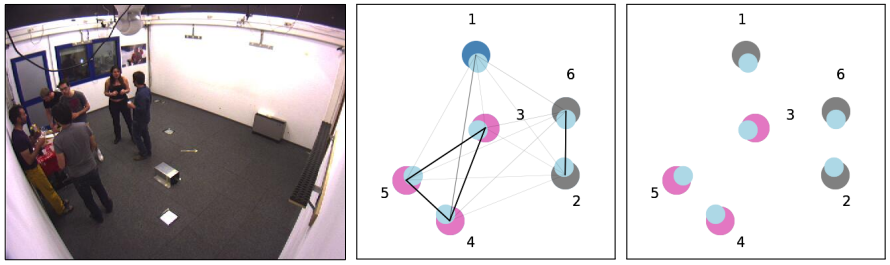


Fig. 5. Example results. *First Column:* Original image from the Cocktail Party dataset, *Second Column:* ground truth conversational group, *Third Column:* results from GCFF [57], *Fourth Column:* our results with DANTE. The wall with the door in the images corresponds to the **top** side of the diagrams in the second to fourth columns. People's colors indicate groups, and line thickness indicates DANTE's affinity prediction (thicker means close to 1 in the [0,1] range). In comparison to DANTE, GCFF tends to be more inclusive, resulting in incorrect large groups. This result aligns with prior findings [65].

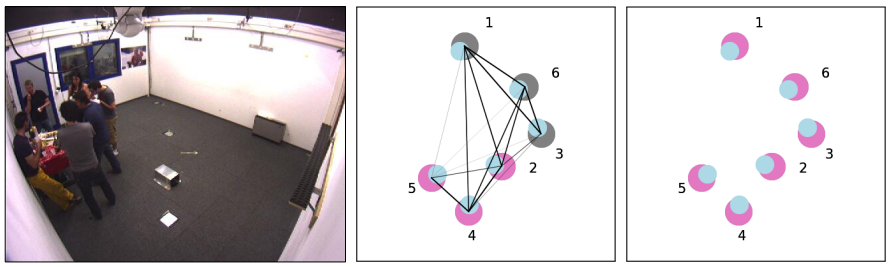
terms of F1 score, it underperforms in Fold 1, as can be seen in the Appendix (Table 5). One reason for this discrepancy is that Fold 1 had the most noisy spatial features $\mathbf{f}$ for the individuals in the scene. This hurt prior work, but was especially harmful to our data-driven method.

4.6 Qualitative Results

Figures 5 and 6 provide qualitative results in the Cocktail Party dataset. In particular, Fig. 5 compares example results between our method and the GCFF approach [57]. In rows 1, 3, and 4, GCFF chooses larger groups due to a penalty on small group sizes. This preference for larger groups often overrides information in the data, such as a person facing away from the proposed group. In row 2, Person 5 is likely excluded from GCFF's primary group due to a heuristic in the GCFF algorithm which prevents grouping two people if there is someone else in-between them. In comparison, our deep learning approach considers the social context of the groups to learn more nuanced spatial patterns, such as how one orients oneself when leaving a group (row 1), how one behaves when standing



(a) DANTE estimates that the person 1 does not belong to a conversational group while he is grabs an object from the coffee table. Ground truth groups: $g_1 = \{1, 3, 4, 5\}, g_2 = \{2, 6\}$.



(b) DANTE estimates two groups while all individuals are conversing together. Ground truth groups: $g_1 = \{1, 2, 3, 4, 5, 6\}$.

Fig. 6. Failure cases by the proposed approach. *Left:* original image from the Cocktail Party dataset, *Middle:* estimated groups by DANTE, *Right:* ground truth. Individuals are colored based on their groups in each case.

on the outskirts (rows 2 and 3), and how people arranged in a ring can still form smaller groups (row 4). This flexibility largely comes from not employing brittle heuristics to account for context and instead allowing the model to learn from data. Also note that the edge-weights are either very correctly thick due to high predicted affinities (> .9) or correctly thin (< .2) in these examples. In general, we found DANTE to be very confident when making successful predictions.

Figure 6 shows failure cases by our method. In Fig. 6(a), one of the main limitations of our approach becomes evident: useful information (e.g. posture or gaze) to correctly assign the individual 1 to the right group is not available to DANTE using the default spatial features described in Sec. 3.2.4. Our method only has access to 2D position and orientation in the Cocktail Party dataset, which can make some interaction analysis difficult. In case (b), another limitation is apparent: DANTE lacks environmental features (e.g. table or wall locations), which could explain the large space in between the two predicted groups. Instead, DANTE infers this large empty space to signify that the two cohorts are separate conversations. Particularly in case (b), DANTE predicts many intermediate affinity values with inconsistent weights across some node clusters, an indication of the uncertainty in the affinity predictions and uncertainty in determining the group structures. These failure cases illustrate opportunities for future improvement, as further discussed in Section 7.

4.7 Analysis of the Effect of the Context Transform in DANTE

We hypothesized that adding contextual information to the affinity computation by DANTE would improve the results of our group detection approach. To explore if this was effectively the case, we performed a small ablation study. In particular, we evaluated a version of DANTE that only reasoned about the position and orientation of the individuals of interest using the Dyad Transform.

The results for the Cocktail Party dataset are presented in Table 1 in the row corresponding to DANTE-NoContext. As expected, excluding the Context Transform from DANTE resulted in 9% worse average group detection performance, although a t-test on the difference in F1 scores between DANTE and DANTE-NoContext was not significant (p=0.39). We attribute the lack of significance to results being similar in two out of the five folds (see Table 3 in the Appendix).

Figure 7 shows example, qualitative results in the Cocktail Party dataset. Comparing DANTE vs. DANTE-NoContext, we can observe that the social context input to the affinity computation is highly relevant to the group detection task. In Fig. 7(a), two people that are separated by another interaction are grouped with one another, even though this would be unlikely in real situations. In real life, the two people would have trouble communicating with each other when a conversation is happening in-between them. Although it may seem surprising that persons 1 and 4 could be given high affinity, there were large groups in the datasets with members at similar distances. The context of the other people in the room thus became necessary to differentiate these cases. In Fig. 7(b), a person is missed in a group interaction, due to slightly lower affinities than the rest of the pairwise interactions. Overall, we see that without context, DANTE is often unsure of what to do with people at greater distances, and these less confident affinities can lead to incorrect groupings. This suggests that our complete version of DANTE is better at reasoning about complex spatial patterns.

The results in the SALSA dataset were similar to the Cocktail Party. We found that removing the Context Transform from DANTE led to reduced average F1 scores across all folds (0.65 for DANTE vs. 0.57 for DANTE-NoContext). Although DANTE was consistently better with the context information, a t-test resulted in no significant differences (p=0.33).

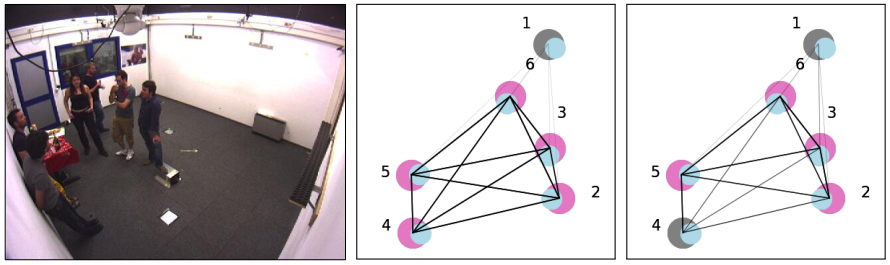
In Coffee Break, the average group detection results were slightly better without the Context Transform (Table 1). Although the difference was not significant (p=0.86), it surprised us. We attribute the lack of benefit of the context in this case to the more noisy annotations and less data provided by this dataset. This idea is supported by the fact that DANTE performs comparatively worse than DANTE-NoContext on the noisiest fold (Fold 1 in Table 5 within the Appendix). We believe that this is due to the increased complexity of DANTE vs DANTE-NoContext, which causes it to overfit on the training data. Worth noting, DANTE-NoContext slightly outperformed other baselines in terms of average F1 score in this benchmark. This finding reinforces the idea that deep learning can help with the group detection task.

5 GENERALIZATION EXPERIMENT

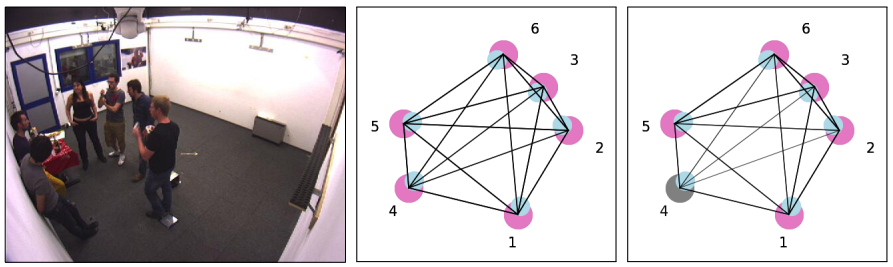
Although our focus is on conversational group detection, we performed an experiment using a large, general group detection dataset. While prior conversational group detection algorithms often relied on heuristics tailored to conversational group detection [31, 57, 62, 65, 67], we hoped that our data-dependent procedure would allow it to perform well outside of its initially intended domain. We followed the same data-augmentation and setup procedures as in Sec. 4 for this experiment.

5.1 Dataset

Friends Meet [5] is composed of 53 synthetic and real sequences of varying group types, including but not restricted to conversational groups. Keeping in line with prior work [61], we restrict our training and evaluation to the synthetic sequences. These sequences were chosen by [61] because the real sequences are not labeled by group type. Also, [61] removed queuing sequences from the data because queues are semantically and spatially different from the other group interactions in the dataset, e.g., groups of pedestrians that walk together towards a destination. Therefore, we present our results based on the 25 non-queuing synthetic sequences, with 200 annotated frames per sequence, for a total of 5,000 frames.



(a) Without global context, DANTE-NoContext groups people 1 and 4, even though they are clearly occluded. Ground truth groups: $g_1 = \{2, 3, 4, 5, 6\}, g_2 = \{1\}$ (not interacting).



(b) DANTE accounts for the large group when computing pairwise-affinities, while DANTE-NoContext gives Person 4 lower pairwise-affinities based on distance and an inability to notice the large group. Ground truth groups: $g_1 = \{1, 2, 3, 4, 5, 6\}$.

Fig. 7. Ablative analysis. *Left:* Image from the Cocktail Party dataset, *Middle:* DANTE, *Right:* DANTE-NoContext. Grouped individuals share the same color.

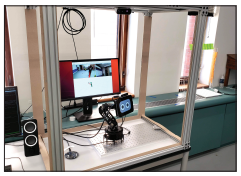
Because Friends Meet only provides position information for potential interactants, we perform the same feature augmentation as [61], which subtracts people's position between consecutive frames to obtain crude velocity estimates. These estimates are normalized and used in place of the orientation features considered in DANTE's default input representation (Sec. 3.2.4). Using the instantaneous velocity instead of the orientation can be thought of as assuming that people orient in the motion direction. Although this assumption breaks down when individuals are standing still.

5.2 Evaluation Metrics

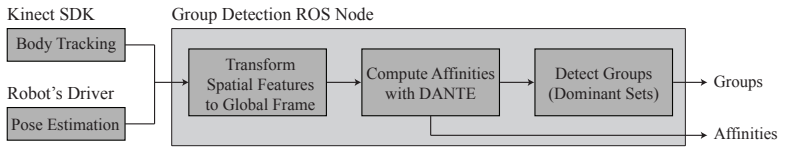
We consider both the challenging $T = 1$ F1 metric, as well as the more lenient Group Detection Success Rate (GDSR) employed in prior work [61]. GDSR measures the percentage of groups correctly identified, and a group is considered correct if at least 60% of its members are detected.

5.3 Results

The results are presented in Table 2. Our approach (DANTE row) achieved state-of-the-art results on this general group detection benchmark [5]. As in Sec. 4, we conducted independent t-tests to compare methods in terms of their F1 and GDSR scores across folds. In comparison to GTGC [62], our approach led to significantly higher F1 scores ($t(6.05)=-9.79$, $p<0.001$) as well as higher GDSR, ($t(4.12)=-5.71$, $p=0.004$). There was also a significant difference between the F1 scores of GCFF [57] and our approach, ($t(4.82)=2.91$, $p=0.03$). Lastly, the difference in terms of GDSR for the latter two methods was close to significant ($p=0.08$).



(a) Robot Setup



(b) Group detection system integrated with the Robot Operating System (ROS)

Fig. 8. *Left:* Robot setup. *Right:* Components of our interactive system.

6 APPLICATIONS

Our group detection approach can be used to increase the social awareness of interactive systems. To demonstrate this in practice, we built an interactive system using the Robot Operating System (ROS) [51], a popular collection of tools and libraries for robotics development.[2] This effort adds to a larger body of work that shows that automatic group detection is relevant in HCI, ranging from mobile systems [41] to robotic interfaces [20, 33].

6.1 Group Detection for Human-Robot Interaction

In our demonstration application, a table-top robot is used to identify F-Formations based on users' spatial behavior relative to each other and its own spatial configuration in our lab environment (Figure 8). The main components of our interactive system are a robot arm with a screen face, and two RGB-D cameras (Fig. 8a). The robot and all sensors are connected to a nearby desktop computer, which processes data in real-time and controls the robot. The computer has an NVidia GeForce GTX 1080 Ti graphics card that we use for testing DANTE.

6.1.1 Scenario. The robot, Shutter, acts as a photographer in our demonstration. The robot uses a forward facing RGB-D camera to detect faces such that when an individual or a group approaches the robot, it turns towards them and asks them if they would like to be photographed. If they accept, Shutter then counts down from 3 and takes a photo. While Shutter interacts with people, we use the secondary RGB-D camera that is fixed above the robot to reason about groups. This camera is a Kinect Azure. It provides a wider view of the environment, augmenting the robot's own sensors.

6.1.2 Spatial Features. As illustrated in Fig. 8b, we use the fixed camera above the robot for human body tracking in real-time via the Azure Kinect Body Tracking SDK [44]. The Kinect SDK provides keypoint tracking information for each person in view. This information includes position and head orientation, which we use for spatial reasoning.

We consider the robot as a social agent in the demonstration scenario because it socially engages in conversations with users. For the robot's spatial features, we use its position on the table, which is known a priori and fixed based on our setup. For its head orientation, we use the orientation of its screen face, which is obtained from its joint positions.

[2]We will open-source our ROS wrapper along with our implementation of DANTE to facilitate future replication efforts.

Table 2. Generalization Results on Friends Meet Dataset

Model	$T = 1$	GDSR
GTCG [62]	0.66	0.90
GCFF [57]	0.79	0.95
DANTE	**0.90**	**0.99**

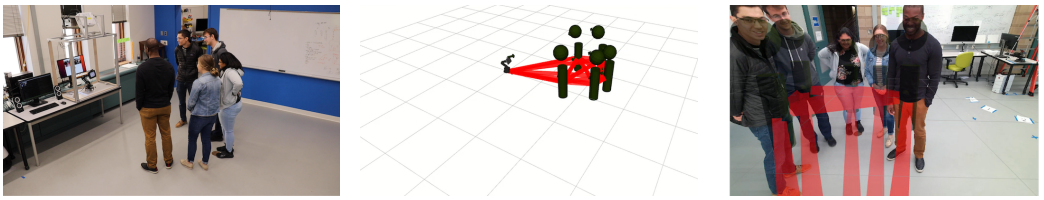


Fig. 9. Shutter interacting with a group after taking a photo and talking to users. All markers are colored dark green in the middle image, indicating that all agents belong to one group. Best viewed in digital form.

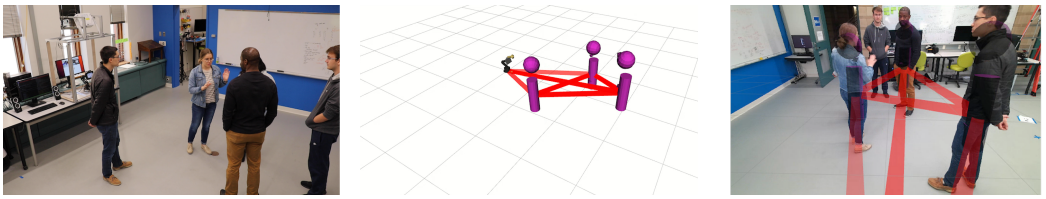


Fig. 10. Example in which Shutter is excluded from a human group (the robot's marker is yellow in the middle image, while the people are purple). Best viewed in digital form.

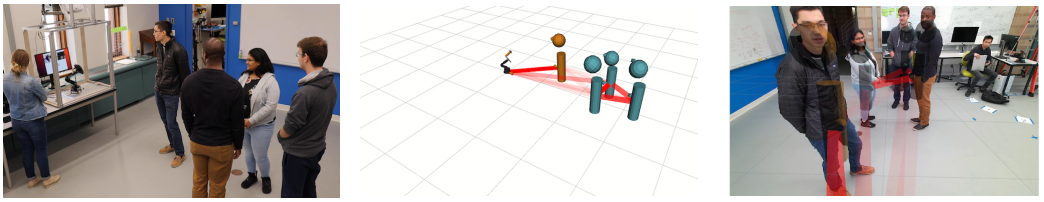


Fig. 11. Example of Shutter interacting with a single individual while a separate group talks in the background. Shutter's marker is orange - the same color as the person closest to the camera. Best viewed in digital form.

6.1.3 Real-Time Group Detection. All the group detection logic is encapsulated into ROS node, a computer process that runs within the ROS network of the robot (Figure 8b). At a given time, the group detection node processes spatial information as follows. First, it transforms the latest vision-based person tracking information and the robot's pose information that has been received into a shared, global coordinate frame. The positions and orientations are then organized into an interaction graph, so that affinities can be computed for all pairs of social agents using DANTE (pre-trained on the Cocktail Party dataset used for the evaluation of Sec. 4). Then, the affinities are processed by the Dominant Sets algorithm, which finally outputs groups. Note that our system also outputs affinity scores for visualization purposes, as illustrated in Figures 10 and 11. In these Figures, the affinity values are displayed as red lines between the social agents in the scene. The opacity of the lines indicates the strength of the affinities (more opaque means stronger). In the middle and right images of Figures 10 and 11, each detected person is denoted with a cylindrical marker, while Shutter is depicted as a 3D model of the robotic arm. The people markers are colored according to the group to which they belong. That is, people predicted to be within the same group are of the same color. Shutter also has a colored arrow above it that indicates its group affiliation.

6.1.4 Qualitative Results. We investigate the performance of DANTE under three scenarios: groups including Shutter, groups excluding Shutter, and multiple groups within the same frame. Figure 9

shows an example of the first case, in which Shutter interacts with a group of five individuals. The pairwise affinities between each social agent are strong, corresponding to opaque lines connecting the agents. Overall, our method worked well in these types of scenarios, except when Shutter moved its face to one side and suddenly the group divided in two. Sometimes the split was correct. The robot was trying to interact with a single individual, whose face was detected by its camera. But sometimes the motion was due to the robot turning around during part of the interaction to signal to users that they should look at the screen behind the robot to see their picture. Our group detection system had no information about this change in activity, but this information could be considered by it in the future.

After Shutter took a photo, the group moved back and formed a group distinct from Shutter (Figure 10). In this scenario, Shutter is detected to be in a separate group from the other people in the room. But because of the setup of the Kinect camera, the other two people in the group were occluded and their poses were not detected. Thus, they were not grouped together in this case.

In general, the algorithm detected groups distinct from Shutter's, but occasionally had problems in separating the social agents standing in close proximity to the robot. This sometimes happened because the robot was looking toward another group, or because members of the group looked at the robot over their shoulders or left space for Shutter to look into the group. Considering additional spatial features, such as body and shoulder orientation, during group detection could help reduce the number of errors that occur in these situations.

Our approach is also capable of detecting multiple groups in real-time (often running at 15 Hz, the framerate of the Kinect camera). For example, consider Figure 11. In this case, one person interacts with Shutter and there is another group in the background. Note that the affinities between the group in the foreground and the background are nonzero, but significantly weaker than the affinities between group members. When DANTE misclassified the groups, it was often because the members of one group were looking toward the other, as described before. But overall, the algorithm performed well at separating distinct groups that occurred simultaneously.

6.2 Other Applications

We envision using our group detection approach in other applications that benefit from social intelligence. For example, our approach could be used to improve social robot navigation, including delivery and guide robots [21, 60], autonomous cars [55], and assistive wheelchairs [26]. Additionally, our method could be used to create better embodied conversational assistants in office environments [6, 7] or educational settings [43]. Finally, we look forward to testing our approach for creating interactive public installations, like facades that respond to the group activities of nearby pedestrians [18]. All these types of interfaces require increased levels of social intelligence, which we believe our approach can contribute to. Naturally, the performance of our method will be subject to the availability of appropriate training data, even if it is borrowed from another related domain as in our example demonstration.

7 DISCUSSION

7.1 Evaluation

In comparing our method's performance against prior work across datasets, we observed that our model tended to perform better on larger, better annotated datasets. This is promising because reliable spatial features are quickly becoming a commodity, e.g., as provided by the Kinect Azure SDK that we used in our example application.

We expect data-driven approaches, like ours, to further improve in the future as more data becomes available for group detection. For the time being, our evaluation with small datasets

showed that our approach performed better or as well as prior DL approaches [54, 58] across datasets. Additionally, our approach outperformed baselines on the Friends Meet dataset, a more general group detection benchmark than those considered in Sec. 4. We believe that the success of our approach was in part due to its data-driven nature, as well as thanks to its structure. We predicted affinity values for dyads instead of directly aiming to find groups, e.g., as in [58]. This choice effectively increased the number of examples that DANTE had to learn from by a factor of 15 to 153, depending on the number of people in the scene, which helped avoid overfitting. Our results also suggest that DANTE effectively processed spatial data from a variable number of unordered social agents. Thus, our work reinforces the idea that symmetric functions are an effective mechanism for neural networks to deal with input sets [28, 50].

7.2 Improving Social Awareness in Practice

We showed the applicability of our group detection approach in a human-robot interaction scenario. In this demonstration, our approach ran in real-time in a standard desktop machine. Because we considered the robot as one more social agent, our method detected not only human groups, but also human-robot groups. Furthermore, our approach was able to detect when people joined or leaved the robot's conversation. The availability of this information opened up possibilities to improve the interaction with our robot photographer, e.g., by enabling it to better frame photos, or simply acknowledging the changes in its social context, which can help queue users of the photography system.

Because DANTE is easily extensible, an interesting avenue for future work would be to augment the spatial features considered by our method. For example, other relevant features include: additional spatial features for the social agents, like body pose, which can signal social information [34]; features that describe the layout of the environment, which can affect F-Formations [36]; or even information about the type of agent being analyzed. We believe that the latter type of information will be relevant for heterogeneous social interactions, e.g., interactions among people and virtual agents or robots. In these situations, agents may communicate attention and engagement through different non-verbal social signals. The relevance of these signals could potentially be captured by data-driven methods, like ours, to further improve group detection.

7.3 Limitations

Our work is not without limitations. First, our method does not take advantage of the temporal correlation of spatial features captured by a situated sensor in the world. But this type of information could potentially improve group detection [65]. Second, our method's runtime is also an important consideration. DANTE scales quadratically with the number of participants in the scene and the Dominant Sets [31] clustering algorithm scales cubically. While this issue did not prevent us from performing real time group detection in our demonstration, it could be an issue if applied to more crowded settings. Third, our demonstration of our group detection approach was conducted in a laboratory setting. Further experiments are needed to validate the robustness of our approach in more dynamic scenarios.

8 CONCLUSIONS

We presented a novel approach for conversational group detection. Our method combined graph clustering with modern deep learning techniques to identify group interactions based on visual patterns of spatial behavior. Under the challenging T=1 F1 metric, our method significantly outperformed or performed as well as previous methods in a variety of conversational group detection benchmarks. Additionally, we obtained good results under the GDSR metric in a more general group detection task, showing the generalization capabilities of our proposed approach. From an

algorithmic point of view, clear improvements were derived from better affinity scores used for graph clustering in comparison to prior work. Additionally, the use of data-driven methods allowed our approach to cope with complex spatial patterns of behavior without ad-hoc steps to verify group interactions. These features made our approach robust and practical to be applied in a real human-robot interaction scenario.

REFERENCES

[1] Jake K Aggarwal and Michael S Ryoo. 2011. Human activity analysis: A review. *ACM Computing Surveys (CSUR)* 43, 3 (2011), 16.

[2] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe. 2016. SALSA: A Novel Dataset for Multimodal Group Behavior Analysis. *IEEE PAMI* 38, 8 (Aug 2016), 1707–1720. https://doi.org/10.1109/TPAMI.2015.2496269

[3] Stefano Alletto, Giuseppe Serra, Simone Calderara, Francesco Solera, and Rita Cucchiara. 2014. From ego to nos-vision: Detecting social relationships in first-person views. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 580–585.

[4] Till Ballendat, Nicolai Marquardt, and Saul Greenberg. 2010. Proxemic interaction: designing for a proximity and orientation-aware environment. In *ACM International Conference on Interactive Tabletops and Surfaces*. ACM, 121–130.

[5] L. Bazzani, M. Cristani, and V. Murino. 2012. Decentralized particle filter for joint individual-group tracking. In *CVPR*.

[6] Dan Bohus, Sean Andrist, and Eric Horvitz. 2017. A study in scene shaping: Adjusting F-formations in the wild. In *Proceedings of the 2017 AAAI Fall Symposium: Natural Communication for Human-Robot Collaboration*.

[7] Dan Bohus and Eric Horvitz. 2009. Dialog in the open world: platform and applications. In *Proceedings of the 2009 international conference on Multimodal interfaces*. ACM, 31–38.

[8] Dan Bohus and Eric Horvitz. 2009. Learning to predict engagement with a spoken dialog system in open-world settings. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 244–252.

[9] Dan Bohus, Chit Saw, and Eric Horvitz. 2014. Directions Robot: In-the-wild experiences and lessons learned. In *AAMAS*.

[10] Oliver Brdiczka, Jérôme Maisonnasse, and Patrick Reignier. 2005. Automatic detection of interaction groups. In *Proceedings of the 7th international conference on Multimodal interfaces*. ACM, 32–36.

[11] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

[12] Ming-Ching Chang, Nils Krahnstoever, and Weina Ge. 2011. Probabilistic group-level motion analysis and scenario recognition. In *Proc. of the 2011 International Conference on Computer Vision (ICCV)*. 747–754.

[13] Chih-Wei Chen, Rodrigo Cilla Ugarte, Chen Wu, and Hamid Aghajan. 2011. Discovering social interactions in real work environments. In *Proc. of Face and Gesture 2011*. 933–938.

[14] Wongun Choi, Khuram Shahid, and Silvio Savarese. 2009. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *Proc. of the 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)*. 1282–1289.

[15] Tanzeem Choudhury and Alex Pentland. 2002. The sociometer: A wearable device for understanding human networks. In *CSCW'02 Workshop: Ad hoc Communications and Collaboration in Ubiquitous Computing Environments*.

[16] 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.

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

[18] Peter Dalsgaard and Kim Halskov. 2010. Designing urban media façades: cases and challenges. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2277–2286.

[19] Elwys De Stefani and Lorenza Mondada. 2014. Reorganizing mobile formations: When âĂIJguidedâĂİ participants initiate reorientations in guided tours. *Space and Culture* 17, 2 (2014), 157–175.

[20] Eyal Dim and Tsvi Kuflik. 2015. Automatic detection of social behavior of museum visitor pairs. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 4, 4 (2015), 17.

[21] Vanessa Evers, Nuno Menezes, Luis Merino, Dariu Gavrila, Fernando Nabais, Maja Pantic, and Paulo Alvito. 2014. The development and real-world application of frog, the fun robotic outdoor guide. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 281–284.

[22] Alircza Fathi, Jessica K Hodgins, and James M Rehg. 2012. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1226–1233.

[23] 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*. ACM, 937–946.

[24] Weina Ge, Robert T Collins, and Barry Ruback. 2009. Automatically detecting the small group structure of a crowd. In *Proc. of the 2009 Workshop on Applications of Computer Vision (WACV)*. IEEE, 1–8.

[25] Erving Goffman. 2008. *Behavior in public places*. Simon and Schuster.

[26] Isabella Gomez Torres, Gaurav Parmar, Sammarth Aggarwal, Nathaniel Mansur, and Alec Guthrie. 2019. Affordable Smart Wheelchair. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, SRC07.

[27] Georg Groh, Alexander Lehmann, Jonas Reimers, Marc René Frieß, and Loren Schwarz. 2010. Detecting social situations from interaction geometry. In *Proc. of the 2010 IEEE Second International Conference on Social Computing*. IEEE, 1–8.

[28] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2255–2264.

[29] Edward Twitchell Hall. 1910. *The hidden dimension*. Vol. 609. Garden City, NY: Doubleday.

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

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

[32] Helge Hüttenrauch, Kerstin Severinson Eklundh, Anders Green, and Elin A Topp. 2006. Investigating spatial relationships in human-robot interaction. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 5052–5059.

[33] Junko Ichino, Kazuo Isoda, Tetsuya Ueda, and Reimi Satoh. 2016. Effects of the display angle on social behaviors of the people around the display: A field study at a museum. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 26–37.

[34] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. 2019. Towards Social Artificial Intelligence: Nonverbal Social Signal Prediction in A Triadic Interaction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10873–10883.

[35] Manuela Jungmann, Richard Cox, and Geraldine Fitzpatrick. 2014. Spatial play effects in a tangible game with an f-formation of multiple players. In *Proceedings of the Fifteenth Australasian User Interface Conference-Volume 150*. Australian Computer Society, Inc., 57–66.

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

[37] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014).

[38] Hideaki Kuzuoka, Yuya Suzuki, Jun Yamashita, and Keiichi Yamazaki. 2010. Reconfiguring spatial formation arrangement by robot body orientation. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 285–292.

[39] Oswald Lanz. 2006. Approximate bayesian multibody tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 9 (2006), 1436–1449.

[40] Nicolai Marquardt, Robert Diaz-Marino, Sebastian Boring, and Saul Greenberg. 2011. The proximity toolkit: prototyping proxemic interactions in ubiquitous computing ecologies. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 315–326.

[41] Nicolai Marquardt, Ken Hinckley, and Saul Greenberg. 2012. Cross-device interaction via micro-mobility and f-formations. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 13–22.

[42] Paul Marshall, Yvonne Rogers, and Nadia Pantidi. 2011. Using F-formations to analyse spatial patterns of interaction in physical environments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, 445–454.

[43] Yoichi Matsuyama, Arjun Bhardwaj, Ran Zhao, Oscar Romeo, Sushma Akoju, and Justine Cassell. 2016. Socially-aware animated intelligent personal assistant agent. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*. 224–227.

[44] Microsoft. 2019. Azure Kinect SDK (K4A). https://github.com/microsoft/Azure-Kinect-Sensor-SDK. [Online; accessed 14-October-2019].

[45] Alejandro Moreno, Robby van Delden, Ronald Poppe, and Dennis Reidsma. 2013. Socially aware interactive playgrounds. *IEEE pervasive computing* 12, 3 (2013), 40–47.

[46] Daniel Olguín Olguín, Benjamin N Waber, Taemie Kim, Akshay Mohan, Koji Ara, and Alex Pentland. 2009. Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 1 (2009), 43–55.

[47] Hyun S Park, Eakta Jain, and Yaser Sheikh. 2012. 3d social saliency from head-mounted cameras. In *Advances in Neural Information Processing Systems*. 422–430.

[48] Massimiliano Pavan and Marcello Pelillo. 2007. Dominant sets and pairwise clustering. *IEEE transactions on pattern analysis and machine intelligence* 29, 1 (2007), 167–172.
[49] Ashwini Pokle, Roberto Martín-Martín, Patrick Goebel, Vincent Chow, Hans M Ewald, Junwei Yang, Zhenkai Wang, Amir Sadeghian, Dorsa Sadigh, Silvio Savarese, et al. 2019. Deep Local Trajectory Replanning and Control for Robot Navigation. *arXiv preprint arXiv:1905.05279* (2019).
[50] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 652–660.
[51] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: an open-source Robot Operating System. In *ICRA workshop on open source software*, Vol. 3. Kobe, Japan, 5.
[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] 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.
[54] Navyata Sanghvi, Ryo Yonetani, and Kris Kitani. 2018. Learning Group Communication from Demonstration. In *Workshop on Models and Representations for Natural Human-Robot Communication at the 2018 Robotics: Science and Systems Conference (RSS)*.
[55] Friederike Schneemann and Patrick Heinemann. 2016. Context-based detection of pedestrian crossing intention for autonomous driving in urban environments. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2243–2248.
[56] 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, 3547–3551.
[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] Mason Swofford, John Peruzzi, and Marynel Vázquez. 2018. Conversational Group Detection With Deep Convolutional Networks. *arXiv e-prints*, Article arXiv:1810.04039 (Oct 2018), arXiv:1810.04039 pages. arXiv:cs.CV/1810.04039
[59] Lili Tong, Audrey Serna, Simon Pageaud, Sébastien George, and Aurélien Tabard. 2016. It's not how you stand, it's how you move: F-formations and collaboration dynamics in a mobile learning game. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 318–329.
[60] Rudolph Triebel, Kai Arras, Rachid Alami, Lucas Beyer, Stefan Breuers, Raja Chatila, Mohamed Chetouani, Daniel Cremers, Vanessa Evers, Michelangelo Fiore, et al. 2016. Spencer: A socially aware service robot for passenger guidance and help in busy airports. In *Field and service robotics*. Springer, 607–622.
[61] Sebastiano Vascon and Loris Bazzani. 2017. Chapter 3 - Group Detection and Tracking Using Sociological Features. In *Group and Crowd Behavior for Computer Vision*. Academic Press. https://doi.org/10.1016/B978-0-12-809276-7.00004-7
[62] 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.
[63] Marynel Vázquez. 2017. *Reasoning About Spatial Patterns of Human Behavior During Group Conversations with Robots*. Ph.D. Dissertation. The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.
[64] 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 *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 42–52.
[65] Marynel Vázquez, Aaron Steinfeld, and Scott E Hudson. 2015. Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 3010–3017.
[66] Danny Wyatt, Tanzeem Choudhury, and Jeff Bilmes. 2007. Conversation detection and speaker segmentation in privacy-sensitive situated speech data. In *Eighth Annual Conference of the International Speech Communication Association*.
[67] Ting Yu, Ser-Nam Lim, Kedar Patwardhan, and Nils Krahnstoever. 2009. Monitoring, recognizing and discovering social networks. In *Proc. of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1462–1469.
[68] Gloria Zen, Bruno Lepri, Elisa Ricci, and Oswald Lanz. 2010. Space speaks: towards socially and personality aware visual surveillance. In *Proc. of the 1st ACM International Workshop on Multimodal Pervasive Video Analysis*. 37–42.

A APPENDIX

Due to limited space, the evaluation described in the paper only presented overall results (averaged across five folds) for common group detection benchmarks. To supplement these results, Section

Table 3. F1 results (T=1) for the Cocktail Party dataset.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Overall
GComm	-	-	-	-	-	0.60
GTCG	0.31	0.46	0.09	0.17	0.43	0.29
GCFF	0.49	0.68	0.52	0.71	0.80	0.64
DANTE	**0.73**	0.80	0.56	0.72	**0.83**	**0.73**
DANTE+Synthetic	0.70	**0.83**	**0.59**	0.64	0.79	0.71
DANTE-NoContext	0.60	0.68	0.33	**0.75**	**0.83**	0.64

Table 4. F1 results (T=1) for the SALSA dataset.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Overall
GTCG	0.39	0.44	0.41	0.47	0.50	0.44
GCFF	0.59	0.40	0.25	0.24	0.58	0.41
DANTE	**0.76**	**0.57**	**0.70**	**0.50**	**0.69**	**0.65**
DANTE+Synthetic	0.74	**0.57**	0.59	**0.50**	**0.69**	0.62
DANTE-NoContext	0.70	0.41	0.63	0.46	0.64	0.57

Table 5. F1 results (T=1) for the Coffee Break dataset.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Overall
GComm	-	-	-	-	-	0.63
GTCG	0.48	0.32	0.51	0.48	0.58	0.48
GCFF	**0.50**	0.35	0.73	0.71	0.84	0.63
DANTE	0.39	0.40	0.74	**0.77**	**0.89**	0.64
DANTE+Synthetic	0.48	0.36	**0.77**	**0.77**	**0.89**	0.65
DANTE-NoContext	0.46	**0.43**	0.76	**0.77**	**0.89**	**0.66**

A.1 discusses group detection accuracy per fold. Afterwards, Section A.2 discusses additional results when DANTE is trained with the addition of synthetic data, originally generated by Cristani et al. [16]. We hypothesized that synthetic data would help with group detection given the small size of existing datasets. Results did not demonstrate conclusively whether synthetic augmentation improved DANTE.

A.1 Fold-by-Fold Results

Recall that our main focus is to detect conversational groups, as annotated in the Cocktail Party [68], Coffee Break [16], and SALSA [2] datasets. To test the generalization capabilities of our approach to a related but different group detection task, we also provide results for the Friends Meet dataset [5]. Friends Meet was originally created for tracking diverse groups of people.

In Tables 3, 5, 4, and 6 we give a fold-by-fold breakdown of $T = 1$ F1 results for the different datasets, as well as overall (average) scores. Additionally, Table 7 gives a fold-by-fold breakdown of Group Detection Success Rate (GDSR) [61] on the Friends Meet dataset. GDSR measures the percent of groups correctly identified, where a group is considered correct if 60% of its members are included. It is easier than the challenging $T = 1$ F1 metric which requires perfect group detection. Table 7 shows our method nearly saturates the GDSR metric both on a per-Fold and Overall basis.

The fold-level breakdown shows that DANTE often outperforms prior work with consistency and not simply on average. Note that noisy spatial features hurt prior work, but is especially harmful

Table 6. F1 results (T=1) for the Friends Meet dataset.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Overall
GTCG	0.71	0.70	0.60	0.62	0.67	0.65
GCFF	0.85	0.69	0.77	**0.89**	0.76	0.79
DANTE	**0.93**	0.90	0.90	0.86	0.91	0.90
DANTE+Synthetic	**0.93**	**0.91**	0.90	0.87	**0.94**	**0.91**
DANTE-NoContext	0.91	0.89	**0.92**	0.88	0.91	0.90

Table 7. GDSR results for the Friends Meet dataset.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Overall
GTCG	0.95	0.89	0.91	0.86	0.87	0.90
GCFF	0.97	0.89	0.96	0.97	0.97	0.95
DANTE	**0.99**	**0.99**	**0.98**	**0.99**	**0.99**	**0.99**
DANTE+Synthetic	**0.99**	**0.99**	0.96	**0.99**	**0.99**	0.98
DANTE-NoContext	0.98	0.98	0.97	**0.99**	**0.99**	0.98

to our proposed method because of its dependency on data. This noise particularly affected our method in Fold 1 of the Coffee Break dataset.

A.2 Synthetic Data Augmentation

All the tables in the prior page report results from an experiment in which we augment DANTE's training data with synthetic examples from [16]. The synthetic data consists of 100 different situations created by psychologists. In each situation, some simulated people take part in F-Formations and others do not. Ground truth group as well as people positions and orientations are provided as part of this synthetic dataset.

Training DANTE with synthetic data lowered average group detection results by 2% in the Cocktail Party dataset and 3% in the SALSA dataset. In the Coffee Break dataset, the synthetic data slightly improved the results by 1%. We attribute these mixed results to the differences in the amount of data that each benchmark provides and their quality.