Predicting Positions of People in Human-Robot Conversational Groups

Hooman Hedayati
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
hooman@cs.unc.edu

Daniel Szafir
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
daniel.szafir@cs.unc.edu

2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI) | 978-1-6654-0731-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/HRI53351.2022.9889628

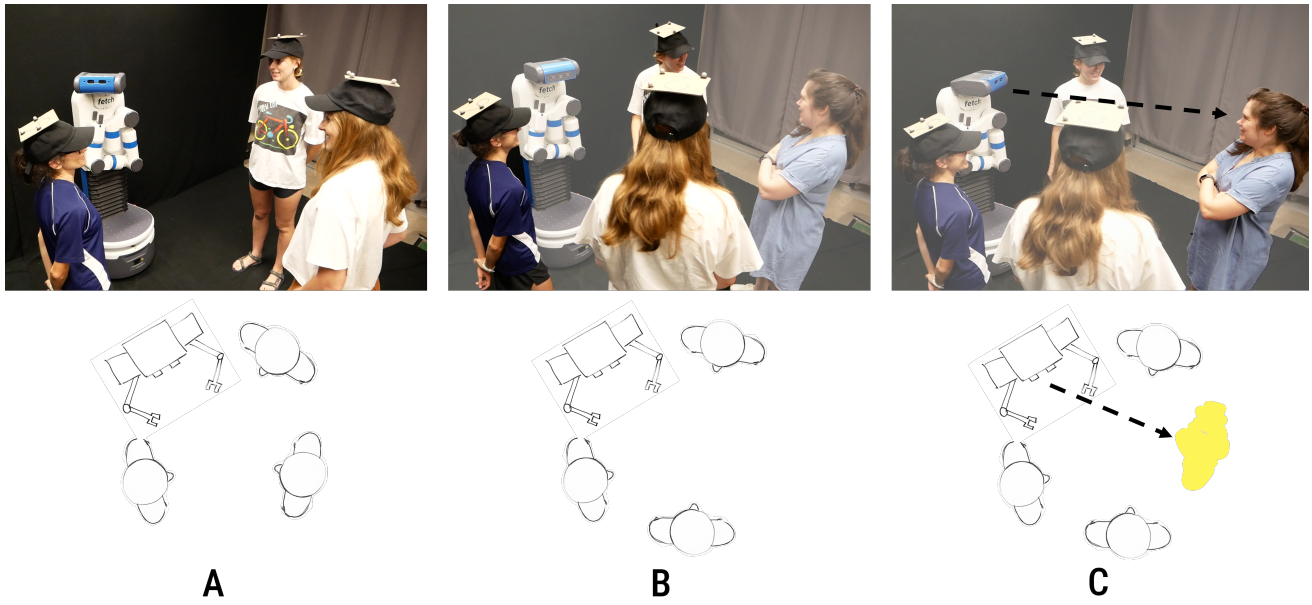

Fig. 1: We present new methods for reasoning about errors resulting in one or more participants being undetected in conversational groups. The top row shows real-world photos, the bottom row shows the robot's perception. A) Three people are in a conversation with a robot and detected (noted with hats). B) A new participant joins the conversation and is not detected. C) Our algorithms predict both the number of missing people (1 in this case) and their likely position, enabling the robot to behave more naturally.

***Abstract*—Robots that operate in social settings must be able to recognize, understand, and reason about human conversational groups (i.e., F-formations). While several algorithms have been developed for identifying such groups, there has been little research on how robots might reason about inaccuracies following group classification (e.g., recognizing only 4 of 5 group members). We address this gap through a data-driven approach that builds knowledge of human group positioning. By analyzing multiple conversational group data sets, we have developed a system for identifying high probability *regions* that indicate areas where people are likely to stand in a group relative to a single *anchor* participant. We use knowledge of these regions to train two models, which we implement on a social robot. The first model can estimate the true size of a partially-observed conversational group (i.e., a group where only some of the participants were detected). Our second model can predict the locations where any undetected participants are likely to reside. Together, these models may improve F-formation detection algorithms by increasing robustness to noisy input data.**

***Index Terms*—F-formations, Social robot, Human-Robot Interaction (HRI), Conversational groups, Data-driven algorithms**

I. Introduction

Robotic technologies are advancing at a rapid pace and are driving the integration of robots into new environments, such as homes and personal spaces, in which they will interact closely with people. For example, social robots are now being used to greet travelers in airports, welcome guests in hotels, and provide directions in shopping malls [1]–[6]. Enabling robots to be aware of how many people are engaged in an interaction with them is a key feature for supporting this transition. Understanding interaction groups can help enable situated spoken language interaction [7], enhance socially-aware navigation in human environments [8], improve natural

behavior generation for robots [9], and sustain user engagement with robots [10].

Being in a conversational group is relatively easy for humans; we are able to position ourselves naturally in conversational groups, acknowledge newcomers, etc. without conscious effort. However, this is challenging for social robots that may interact with one or more people simultaneously due to sensor limitations and the dynamic nature of conversational groups [11]. Many studies have explored the detection of conversational groups by studying proxemics [12], [13], the field of spaces around humans in social settings [14]. Conversational groups are also referred to as F-formations: "An *F-formation* arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access" [15].

F-formations can appear in different arrangements: circular, L, side-by-side, and vis-a-vis. Circular arrangements are formed when there are more than two people in a conversational group, whereas the other three arrangements are most commonly formed when there are only two people in the group. Each arrangement can provide some information about the interaction happening in the group. Kendon [15] describes how L-arrangements happen when two people are in cooperative interactions, whereas vis-a-vis is preferred for competitive interactions. Side-by-side arrangements occur when two people are observing an object, such as a poster on the wall, or when both standing at the edges of a scene against walls [13]. Understanding these F-formation features has been used to improve many human-computer interactions [16]–[20].

Prior work has moved the state-of-the-art toward more accurate F-formation detection [12], [13], [21], [22], but imperfections still remain. There are three challenges in this field. First, input sensors, such as RGB or depth cameras, may produce noisy data due to light conditions or occlusions. As such, the input data may not accurately capture the whole scene. Second, all human detection algorithms contain some amount of error and noise [23], [24]. This may lead to falsely identifying objects in the scene as a human (i.e., false positives), not detecting a human in the scene (false negatives), or inaccurate estimations of head and/or body orientations (as reported in [11]). Finally, F-formation detection algorithms are not fully robust to the dynamics of human groups [12], [13], [21].

These problems can negatively impact interactions when robots are deployed in the real world as social assistants. In particular, this work was motivated by our observation of issues with false negatives in F-formation detection with the Microsoft "directions" robot and the Disney "character" robot. Detection errors for such robots can lead to missed humans in the group feeling ignored or unimportant, thus undermining the central goal of the robot deployment. Moreover, we have found that continuous interaction with groups over time does not often solve misidentifications; instead, detection errors can percolate over time and make misidentifications more likely in the future. As a result, we believe it is essential for social robots to be able to reason about their own confidence in user detection, and ideally correct any errors.

To address this problem and improve the state-of-the-art, we introduce the notion of high probability *regions*. These regions represent the spaces in F-formations in which people tend to occupy. Sec. IV describes how we derive these regions through analyzing two data sets on human conversational groups. In Sec. VI, we introduce a pair of classifiers that demonstrate the utility of understanding such regions, where partial data on F-formation positions can be used to predict the full size of the group (our first classifier) and the likely positions of missing F-formation participants (our second classifier), even when noise is injected into the initial observations. These classifiers may improve the accuracy of prior approaches for detecting F-formations and potentially be used as a sanity check or way of computing classification confidence for existing detection algorithms, which we demonstrate with in-person human robot interactions described in Sec. VIII.

II. Related work

The importance of detecting F-formations has motivated research in how such groups might be identified across a variety of disciplines, including Computer Vision [12], [13], [25]–[27], Human-Computer Interaction (HCI) [21], [28]–[32], Signal Processing & Sensor Fusion [33]–[35], System Engineering [36], [37], Natural Language Processing [38], Robotics [7], [39], [40], and Multi-modal Interaction [10], [41].

At a high-level, such research follows a standard two-step process for identifying conversational groups. The first step is to determine the relative positions and orientations of participants, which can be achieved with a variety of sensors, such as depth cameras, RGB cameras, motion capture cameras, LiDAR, or IMUs. After positions and orientations are acquired, the second step involves feeding this data to an algorithm to reason about the F-formation. This algorithm can be naive, such as using the distances between people, or more complex, utilizing Support Vector Machines (SVM), Graph-Cuts, Hidden Markov Models (HMM), and so on.

Some research builds upon this general method by incorporating additional contextual information about human proxemics, adding additional data processing steps, and/or collecting multimodal data. For example, Brdiczka et al. [28] used HMMs to detect F-formations based on speech activity detection; an automatic speech detector detects which individual stops and starts speaking. Choudhury and Pentland [29] used "sociometer," a custom-built sensor device with a microphone, accelerometer, and IR proxemics sensor, which extracts data for use as an input to a HMM. Hung et al. [42] used a single accelerometer to understand coordinated body movements that can be indicative of being in a conversation. Marquardt et al. [20] used a ceiling-mounted Microsoft Kinect depth sensor to detect ellipses of participants and reason about the F-formation based on the distance and orientation of the participants. Although all these approaches improve F-formation detection, they require additional sensors on participants or augmented environments,which is not feasible in many real-world scenarios.

As an alternative approach, many vision-based algorithms only require a single camera and thus present a more practical

solution for robots. This approach is less invasive, but comes with the cost of increased data noise and may require more complex identification methods. For example, [43] use the Hough-Based tracker to extract participant positions and orientations and a Structural SVM to detect the F-formations, while [44] used Markov Random Fields.

After detecting the positions and orientations of participants, there are several solutions for reasoning about F-formations, such as voting schemes [11], [12], [45], graph-cuts [13], and dominant sets [21], [46]. In recent years, there have also been efforts to use deep learning approaches to detect F-formations automatically [22], [47].

While this research has greatly enhanced our ability to detect and reason about conversational groups, all methods still involve a certain amount of noise and error. For systems that leverage understandings of conversational groups deployed in the real world (e.g., social robots), any classification errors may prove extremely detrimental in terms of both social (e.g., a robot making obvious social miscues or participants feeling ignored) and physical (e.g., a robot positioning itself in an uncomfortable proxemic space or even attempting to navigate through the location of an undetected participant) outcomes. In this work, we introduce an approach that may help remedy such situations by reducing our reliance on the assumption that we have accurately detected all participants in a F-formation. To our knowledge, no prior work has used the position of people as a predictive sanity check for reasoning about F-formation probability. Below, we describe our method for developing such systems that may improve outcomes for any existing F-formation detection algorithm for conversational groups.

III. Approach

The fundamental principle on which this paper is based is our observation that participants in conversational groups tend to occupy certain spatial *regions* within F-formations. We find that these regions adapt with group size and are highly reliable predictors of likely participant locations. Fig. 2 visually illustrates these regions, which are spatial areas defined by probability distributions corresponding to the likely positions of people in a F-formation of a given size. We determined these probability distributions through a data-driven analysis of two open data sets (the SALSA and Babble data sets; for more detail on these data sets see Section V). For example, for F-formations of size 3, we analyzed all size-3 groups across both data sets to identify the most likely positions of people in any size-3 F-formation. This analysis revealed that regions were generally consistent for groups of the same size and varied predictably as group size changed. More details on this process of region identification are found in Section IV.

We can leverage an understanding of such regions to help validate detected F-formations and/or reveal the possibility of errors in a detection algorithm. This knowledge may improve F-formation detection in two principle ways, which we demonstrate across two datasets and a laboratory experiment and implement on a social robot as a systems contribution. First, we can use knowledge of these regions to predict the size of a F-formation when it is partially observed. For example, if an algorithm suggests a potential F-formation, our classifier can independently predict the size of the F-formation in a reliable manner, which may be used as a sanity check for the original F-formation detection algorithm. Second, if there is any inconsistency between a given F-formation and our "F-formation Size Predictor" classifier, the likely location of any people the original algorithm missed can be determined. As a proof-of-concept, we explore a subset of this problem, where the ground truth is that the original F-formation detection algorithm missed one person, for example when a F-formation exists with four individuals (robots or people), but was classified by an algorithm as a F-formation of size three. In such a case, our algorithm can predict that: (1) the original classification of three individuals is incorrect and that the F-formation is more likely to be of size four and (2) where the missing participant is likely to be located. A visual example of such a situation is provided in Fig. 1, where a F-formation prediction algorithm recognized two people as being within a conversational group with the robot, but the third person was not detected; our classifiers can recognize that a person was missed and predict the region where the missing person is likely to be standing.

To accomplish this classification, we first manually assigned discrete class labels to each potential region for various F-formation sizes based on our original region analysis (e.g., region "A," "B," or "C" for a size-3 F-formation). Each class label was defined using the center (x,y) coordinate of the corresponding region probability distribution and a manually specified threshold distance (i.e., radius) from this center. At a high level, our approach in constructing our classifiers was to treat one individual as an *anchor* and evaluate the observed F-formation (which may contain missing data) from the anchor's perspective (i.e., transforming all observations into a coordinate system where the anchor represented the origin). From there, we compare the observed position of participants to the region locations where we would expect participants to be to determine whether we may have missed any individuals (and if so, where they are likely to be). The choice of which individual is treated as an anchor is arbitrary (our system is robust to any decision) and, in our deployments, we use the robot involved in conversational groups as the anchor since observed data is already captured in the robot's frame of reference and thus does not require any extra coordinate transformations. More details on this classification process are found in Section VI.

We evaluated our classifiers on two data sets and in an in-person laboratory experiment with a social robot, examining predictions on true F-formation size and predictions on missing participant location. For size, we found that our classifier could produce accurate predictions of true group size given information on only two participants (i.e., given the positions and orientations of only two individuals, we could predict whether the F-formation was actually a group of size 3, 4, 5, 6, or 7). For participant location, we first validated our approach in simulation by providing the classifier with F-formations where one individual was randomly removed (simulating an algorithm missing a person, possibly due to a sensor error or

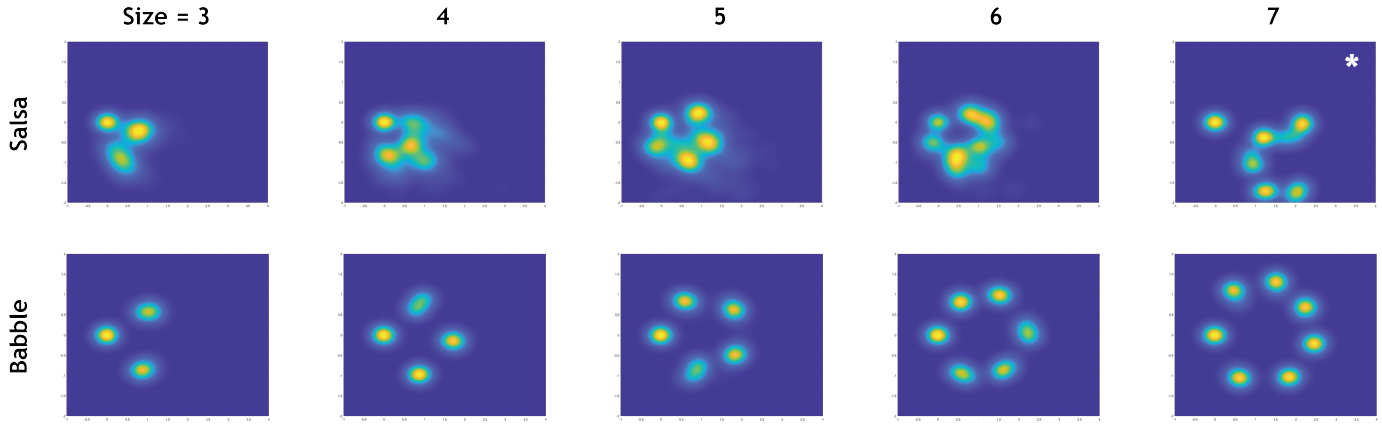

Fig. 2: Heatmaps illustrating probabilistic regions (i.e., participant locations) in different F-formation sizes from the SALSA and Babble data sets. * *indicates where annotation discrepancies occur following manual checks.*

occlusion). The classifier then had to predict the region class label where the missing person was likely located and we again found promising accuracy (between 80–100%). We validated these results by implemented our system on a social robot and conducting a small laboratory experiment, finding that our system resulted in 92.85% accuracy identifying a missing participant in a real-world conversational group.

IV. Region Identification

This section introduces the notation used for the rest of the paper and the operations applied in analyzing data to identify regions. The main approach here is to normalize the data so that F-formations can be compared to one another.

The set of all F-formations of size n in a dataset is F_n. Each set will contain several examples of F-formations of size n, i.e., $F_n = \{f_1, f_2, f_3, ...\}$, where f_i is a snapshot of a F-formation of size n Each instance f_i has n participants, so the set of $f_i = \{P_1, P_2, ..., P_n\}$. We adopt a top-down view of the scene (i.e., an x,y coordinate system) in considering participant poses (positions and orientations). Thus, each participant P_m has three values $P_m = \{x_m, y_m, \theta_m\}$, where x and y are the 2D coordinates for the participant position and θ is the orientation.

To compare various F-formations of equal size, we transform each f_i by choosing an *anchor* participant P_A, selected from all P_n. In essence, once the F-formation is transformed relative to this anchor, it is seen from that individual's perspective. Choosing an anchor P_A and transforming the F-formation with respect to them helps in two ways: (1) it simplifies the process of comparing F-formations to one another and (2) it continues the robot-centric perspective that motivates this work (i.e., in our deployment the robot is the anchor as is commonly used in human-robot scenarios [48]). The anchor can be selected randomly; in training our classifiers, we examined scenes using each individual as a potential anchor. To illustrate, Figure 2 shows regions for each f_i by selecting the participant with the smallest x in the captured frame $P_A = P_m | m \rightarrow Min_{x_m}$. From this anchor, a transformation matrix M_T can be used such that P_A is translated and rotated to $(0, 0, \theta = 0)$ and applied to f_i to determine a new transformed F-formation $f_i^{'} = M_R \times f_i$, where $f_i^{'}$ is the original F-formation transformed with the respect to P_A (i.e., the F-formation as seen from P_A's perspective). Eqn. 1 provides the translation matrix:

$$\begin{bmatrix} x_m^{'} \\ y_m^{'} \end{bmatrix} = \begin{bmatrix} x_m \\ y_m \end{bmatrix} - \begin{bmatrix} x_A \\ y_A \end{bmatrix} \tag{1}$$

If the orientation of the participants in the F-formation dataset is accurate, the rotation matrix $R(-\theta_A)$ is used:

$$R(-\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \ \theta \text{ is the rotation angle} \tag{2}$$

$$f_n^{'} = R_{(-\theta_A)} * f_n \tag{3}$$

If the orientation data is not reliable or there are inconsistencies in the dataset, regions can still be identified and clustered; however, it is not suitable for training classifiers due to the missing feature. To rectify this, additional calculation steps can be used in pre-processing. This can be solved as an optimization problem (Eqn. 4) with all F-formations rotated such that they have the minimum distance and angle from each other. This equation provides θ^*, the new rotation angle for each frame. First, f_i is selected as a referenced F-formation (F_{Ref}). This F-formation can be any arbitrary frame in the F_n set. The only condition is that the F-formation selected for F_{Ref} is manually checked for annotation and ground truth accuracy. Our general reason for proposing θ^* is an approach for calculating group patterns even in a dataset that is noisy/error-prone without having to manually correct many errors.

$$\theta^* = \min_{\theta=-\pi}^{\pi} (Distance(R_{(\theta)} * f_n) - f_{Ref}) \tag{4}$$

$$f_n^{''} = R_{(-\theta^*)} \times f_n^{'} \tag{5}$$

Eqn. 5 takes $f_n^{'}$, the outcome of Eqn. 3, and produces the final product, $F_n^{''}$. This process is applied to all frames to produce the translated set. For the data in this paper, we found that the manual check for F_{Ref} worked well. However, we

note that the selected F_{Ref} could be subjective since it relies on a human annotation. To address this issue, after a run of optimization, a new F_{Ref} can be calculated using the average of each set of F-formations of the same size in $f_n^{''}$, which makes the F_{Ref} selection more robust to noise.

The final processing step applied to the data is to convert from discrete points to a continuous probability space. This can help in two ways: (1) a single point is a poor representation of an individual's body shape so there is a need to represent participants using an area rather than a discrete point and (2) having a normal distribution helps to counter noise and small variations of human movements while they are in a conversational group. We use a 2-D Gaussian function, as shown in Eqn. 6, to fill the adjacent area around the data points. A circular distribution is used to represent the possibility for people to rotate in any direction from the point at which they are standing. An average adult shoulder width is approximately 40 cm [49], so $\sigma = 20cm$ is adopted for the distribution.

$$P(x, y) = e^{-\left(\frac{(x-x_0)^2}{2\sigma^2} + \frac{(y-y_0)^2}{2\sigma^2}\right)} \tag{6}$$

Using this process, we analyzed two open-source data sets to calculate probabilistic regions for F-formations of size 3–7.

V. Datasets

This section describes the two datasets used for identifying regions and training/testing the classifiers in our system. For training and testing, the datasets were randomly split 80:20 into training and testing sets, respectively. We followed standard practice in F-formation detection of splitting F-formation instances into different frames with some frames used in training and others in testing [22], [50]. We present two independent classifiers: one for F-formation size (Sec. VI-A), and another for prediction missing person location (Sec. VI-B).

Many datasets are available for studying F-formations, such as Match & Mingle [51], CoffeeBreak [45], Cocktail [52], and Panoptic [53]. The SALSA [54] and Babble [55] datasets were chosen for exploration in this paper because they both include position and orientation data, include the number of annotated F-formations, and contain a variety of different F-formations of different sizes (e.g., 5 different F-formations of size 3 in Babble). As described in Sec. I, there are many possible variations that an F-formation of size 2 can take and there is a strong correlation between the task and the participant positions for dyadic interactions. As such, for this work, we focus only on F-formations with a size of 3 or greater (i.e., corresponding to a robot and at least two other people).

A. The SALSA Dataset

The SALSA (Synergetic sociAL Scene Analysis) dataset [54] is an open dataset for studying group behavior and social signal processing. SALSA was recorded using four synchronized static RGB cameras (1024x768 resolution) operating at 15 frames per second. In addition to the recording, SALSA contains position, pose, and F-formation annotations for every 3 seconds of data (i.e., one annotation every 45 frames). The authors of SALSA used a dedicated multi-view scene annotation tool to annotate the position, head orientation, and body orientation of each individual. The authors also annotated F-formations, where a F-formation was characterized by the position, head, and body orientations of individuals. It contains data of 18 participants over 60 minutes, with frame annotations every three seconds.

Among the annotated frames, we randomly divided the dataset into a training set (∼80%, corresponding to roughly 320 frames) and a testing set (∼20%, 80 frames).

One limitation of the SALSA dataset (and human annotated datasets in general) is that there are some inconsistencies and errors in the annotations. The SALSA data in particular has a variety of inconsistencies in position and orientation annotations. Since the orientation data is not always accurate, an approximation has to be provided for the orientation angle of P_A and the rest of the F-formation is transformed with respect to that, as described in Sec. IV, Eqn. 4–5. During this process, the ground truth value of θ is not known, but because the relative position of people in the SALSA dataset is known, we can still reason about the SALSA dataset. However, since we cannot know the true θ values, it is impossible to accurately map F-formations from the SALSA coordinate system to the coordinate system of another dataset, such as Babble.

B. The Babble Dataset

The Babble dataset [55] contains highly accurate positions and orientations (measured via a motion-capture system) of six participants playing a social game named "The Resistance" [56]. The game session took approximately 13 minutes including the introduction. The dataset consists of 740 frames (4 images and motion capture camera data for each frame) and contains various F-formation groups that the Babble dataset authors annotated as containing 3, 4, 5, 6 or 7 participants.

VI. Classification

As previously described, noise and other errors can degrade the results of F-formation detection algorithms. This incorrect understanding of a scene may lead to other problems when used in automated systems, for example, a robot may choose a sub-optimal action due to the error. This section outlines a data-driven approach for (1) predicting true F-formation sizes given data about only two participants (an anchor and one other participant) and (2) predicting where a participant missed by a F-formation detection algorithm is likely located. These classifiers may be used to validate and improve the results of existing F-formation detection algorithms by offering a sanity check and additional confidence metrics about the result and can be included in any further reasoning that may be performed by a social robot.

A. Classifier 1: F-formation Size

In this subsection, the goal is to train a classifier that uses the position of a single participant (P_m) relative to an anchor participant (P_A) to predict the size (total number of participants) of a F-formation:

$$Model_{(P_m \in f)} \implies f \in F_n \tag{7}$$

The first step is to create compatible training samples. Each frame that has more than one F-formation per scene is deconstructed into new frames, each containing a single F-formation. For example, a scene in the SALSA dataset with five F-formations would produce five frames for each of the different F-formations in the original scene. Then, permutations with replacement of participants are taken ($Perm^r(n, 2)$) to create pairwise data:

$$\forall i \forall j \in f \rightarrow ((x_i, y_i), (x_j, y_j), Size_f) \tag{8}$$

We ignore all instances of duplicated participants (where $i = j$) as it is not helpful to compare a participant with themselves. From there, we treat the first instance in each pair resulting from Eqn. 8 as the anchor participant P_A. This means that, for example, each frame of a F-formation of size four would produce twelve data points (i.e., $4 \times (4-1)$): three data points when P_0 is treated as the anchor (P_A, P_1), (P_A, P_2), and (P_A, P_3), three where P_1 is the anchor, three where P_2 is the anchor, and three where P_3 is the anchor. As the anchor perspective is adopted consistently (i.e., becomes the origin in the transformed coordinates), (x_i, y_i) can be removed from the terms, resulting in final data with the form $[(x_j, y_j), Size_f)]$, where $Size_f$ encodes the F-formations size and was treated as the class variable we wished to predict. After processing the datasets using this technique, there are 3293 data points from the SALSA dataset and 2109 data points from the Babble dataset for the training. Five different classification models were evaluated, including a variety of Ensemble, KNN, SVM, Naive Bayes, Discriminant, and Tree algorithms.

B. Classifier 2: Missing Region

The previous classifier may be able to identify when a F-formation detection algorithm is inaccurate by independently predicting group size. In such an event (e.g., an existing detection algorithm identifies an F-formation with 4 people, but Classifier 1 predicts the true size to be 5 based on the four who were observed), we next describe an additional classifier that predicts where a missing individual is likely to be located to aid robot decision-making (e.g., a robot could double-check the identified region to see if a participant was occluded and thus missed in the initial detection).

To simplify this problem, in this work we consider only "off-by-one" situations where there is at most a misclassification of one participant between the ground truth and the F-formation detection algorithm output. When a F-formation detection algorithm suggests one more person than Classifier 1 (i.e., a potential false positive), then the identified participant locations can be compared with the regions for P_A and size n that denote the probability distributions for most likely participant locations in order to find the most probable outlier. When there is one person less than suggested by Classifier 1, we take the opposite approach, where we examine regions corresponding to the F-formation size suggested by Classifier 1, map identified participant locations to these regions, and look for the region that lacks a participant. First, all regions are calculated based on the desired F-formation size and the anchor participant position:

$$Regions_{(P_A, size=n)} = \{R_1, R_2, ..., R_{n-1}\} \tag{9}$$

As a reminder, regions R_1, R_2, ... are areas in the coordinate system representing probability distributions for where we would expect participants to be located, relative to P_A for a F-formation of size n. Each region is computed from the aggregate data from all F-formations of size n from only the training portion of the dataset, calculated as a circle whose center coordinates is the $Mean(x, y)$ of all participants recorded in that region and radius is the distance between $Mean(x, y)$ and the furthest participant in the region away from the center (i.e., mathematical representations of the areas shown in Fig. 2).

To train a classifier, for each $f_i \in F_n$ in the training dataset, n samples are created by removing one participant and replacing them with the R_i removed label:

$$[Features, Label] = [(\{P_1, P_2, ..., P_{n-1}\} - \{P_i\}), R_i] \tag{10}$$

For each F-formation of size 3 to 7, classification models were trained such that the order of the features does not matter.

VII. Classification Results

Table I shows the results for the "F-formation Size Classifier" (Sec. VI-A), which seeks to predict the size of the F-formation given the position of only one individual relative to the anchor. For both datasets, the KNN model has the highest accuracy, with 75.9% and 94.3% for SALSA and Babble respectively. This indicates that the KNN model may be used reliably when only knowing the positions of two people in the F-formation. Unfortunately, cross-validation between these two datasets is not possible due to a lack of ground truth orientation data in SALSA, as explained in Sec. V.

The confusion matrices (Fig. 3) provide a more detailed look into the predictions made. Cells along the diagonal show the percentage of data points for which the predicted label is equal to the true label. Other cells indicate where incorrect predictions were made. It can be seen that the predictions are consistently reliable for the Babble dataset. For the SALSA dataset, there is confusion between between F-formations of size 5 and 7 in particular, likely due to the low amount of data for F-formations of size 7 (3 annotated frames).

TABLE I: Accuracy for predicting F-formation sizes given the position of the anchor participant and another participant. Trained and tested on the Babble and SALSA datasets. Holdout validation with 20% held out.

	Accuracy for:	
	Babble	**SALSA**
Fine Tree	92.4%	71.0%
SVM	87.5%	72.1%
Weighted KNN	**94.3%**	**75.9%**
Bagged Trees	93.0%	75.6%

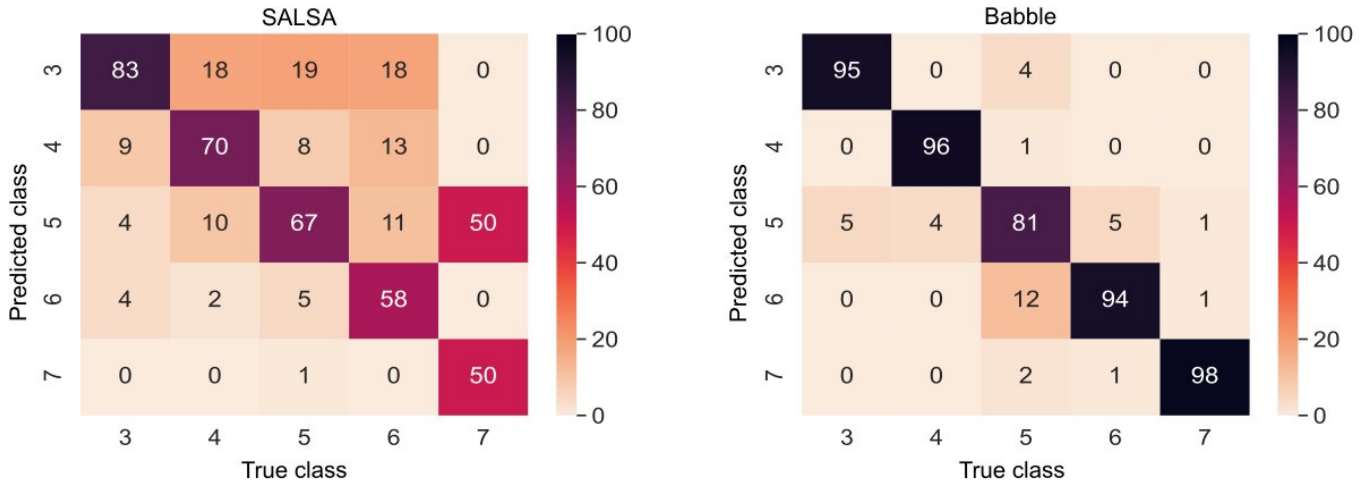

Fig. 3: Confusion matrices for the F-formation size classifier on the SALSA (*left*) and Babble (*right*) datasets. All cell values are percentages. The F-formation sizes form the rows and columns of the matrices. The numbers are in percentages of total testing frames (not the raw number of testing frames, thus they don't add up to the number of frames but do add up to 100).

For the "Missing Region Classifier" (Sec. VI-B), the goal is to predict the position of a missing person when given the correct size of the F-formation (as might be provided by the F-formation Size classifier), but an incorrect set of positions where one participant is omitted entirely (as might be provided by an existing F-formation detection algorithm). Table II shows that the KNN model can predict the position (i.e., region) of a missing person reliably in both datasets; 100% for the Babble dataset and 83% for the SALSA dataset. An accuracy of 100% in the Babble dataset raised concerns about over-fitting. We attribute this accuracy level to the high level of precision regarding measured participant positions and orientations in the Babble dataset and a great degree of similarity of participant positions for F-formations of the same size, as shown in Fig. 2. The results observed for the SALSA dataset may be more representative of typical performance for noisier data or data with more variability in the region positions.

VIII. Laboratory Validation

To better understand how our method may perform in practice, we validated our method in a laboratory experiment. We implemented our classification models in a F-formation reasoning system deployed on a Fetch robot (a mobile ground robot with a manipulator) that interacted socially with participants. Using an IRB-approved protocol, we recruited 8 participants (7 females, 1 male) from the University of Colorado Boulder campus. We conducted a 1-hour laboratory experiment in which we asked participants to have a conversation with the Fetch. First, we asked participants to wear a special baseball hat with reflective markers that enabled high precision tracking within a Vicon motion capture space (i.e., providing ground truth comparisons). This allowed us to track participant head positions and orientations throughout the activity at 100 frames per second with a precision of 1mm. Then, we introduced participants to a social activity in which they had a casual conversation with each other and the robot discussing their last vacation, favorite sports, etc. A researcher acted as a moderator and chose a participant at random to either leave or join the conversational group every 4–5 minutes, enabling us to test our system against groups of different sizes ranging from 3–9 (including the robot) and explore dynamically changing groups.

To evaluate if our system could improve F-formation detection, we created a scenario whereby one of the participants was not tracked by the robot. To accomplish this, when the moderator asked a random participant to leave the group, we had the participant remove their reflective markers so they were no longer visible to the robot. This untracked participant then returned to the conversational group. Figure 1 shows frames where an untracked participant enters the conversation. Out of the whole session, we collected 28 instances of group interactions, comprised of 4 samples from each F-formation size (3–9). For each F-formation size, 2 of the 4 samples consisted of instances where all participants were correctly tracked, while the other 2 instances represented times where one member of the F-formation was not wearing a reflective hat and thus could not be detected by the robot. Including both types of instances helped us test for false-positive (i.e., our system indicating the presence of a missing person even though it had actually not missed any participants) and true negatives (i.e., our system correctly inferring that all participants were currently being tracked and there are no missing participants).

Following our experiment, two members of the research team manually annotated each of group interaction frames. The annotators separately categorized the frames by specifying the F-formation memberships. We compared the annotations for F-formation memberships across the two annotators and found perfect inter-rater reliability (Cohen's kappa) $\kappa = 1$.

Results: For the laboratory validation, we used the KNN classifier trained on 80% of Babble and used the whole pipeline (both classifiers). Overall, we found that our system had an

	Babble Dataset					SALSA Dataset				
	Accuracy for F-formation size:									
	3	4	5	6	7	3	4	5	6	7
Tree	**100.0%**	**100.0%**	99.0%	99.2%	99.6%	86.3%	80.0%	86.0%	86.8%	87.1%
SVM	85.6%	75.0%	**100.0%**	**100.0%**	**100.0%**	**88.3%**	73.0%	88.3%	88.3%	90.6%
Weighted KNN	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**88.8%**	**83.6%**	**88.8%**	**91.4%**	**92.9%**
Bagged Trees	99.6%	98.5%	**100.0%**	**100.0%**	99.6%	87.4%	82.6%	87.4%	88.3%	**92.9%**

TABLE II: Accuracy for predicting regions in Babble & SALSA. Holdout validation with 20% held out.

accuracy of 92.85%, calculated as the correct F-formation predictions divided by the total predictions. In examining our data, we found no instances of false positives, with all errors being false negatives. These mostly occurred in F-formations of size 3. We attribute these errors to the fact that people in a small group size (e.g., 3) have more room to stand compared to groups with more people (e.g., 7), thus regions for small group sizes may have greater variance. Despite these errors, we believe our high overall accuracy rate validates the promise of our approach for use in real social robot interactions.

IX. Discussion & Future Work

Our results showed that the two classifiers (Sec. VI-A & Sec. VI-B) may improve reasoning about partially observed F-formations, both in an analysis on publicly available datasets and when implemented on a social robot that interacted with participants in our laboratory validation. The classifiers can successfully predict the size of a F-formation based on a partially observed scene (i.e., predicting true group size by observing only two participants) and can also predict where a missing person may be located. This could be helpful for social robots that interact with groups of people. We envision these classifiers being added to the end of a social robot's F-formation detection pipeline and note that both classifiers are agnostic to whatever algorithms are used in the other detection steps. Overall, we believe such classifiers may improve the perception that a robot has of the world and might be used to guide various robot behaviors. As an example, if the robot's standard F-formation detection algorithm output identifies an F-formation of size 3, but the output of our F-formation size classifier is 4, the robot could decide to use behaviors that would be suitable for a group of either 3 or 4 participants, it could ask for clarification about the group size from the participants, or it might attempt to use our missing region classifier to determine where a missing participant might be located and take a closer look to determine if indeed a participant was missed in its initial understanding of the group.

While we recognize the importance of comparing new system performance against existing tools, we are unaware of any existing F-formation detection algorithms that can predict missing people, making it difficult to compare our approach to them. Moreover, our goal in this work is to develop systems that can extend, not replace, existing algorithms for F-formation detection (e.g., if an existing detection algorithm matches the output from our size classifier, the robot may have higher confidence that the group was correctly identified). We plan to explore the integration of our classifiers with existing F-formation detection algorithms in future work.

In constructing our second classifier, we focused on a sub-problem of predicting locations for missing participants where only one person was missing from the detected scene. In the future, we intend to explore the reliability of our approach when more than one person is missing. The overall reliability of the probabilistic regions across datasets makes us believe that predicting the locations of multiple missing participants is feasible, although we anticipate increases in uncertainty.

In general, we believe that our data-driven, region-focused approach will be applicable for improving F-formations detection, but acknowledge it may have poor performance on groups with radically different features from those our classifiers were trained on. For instance, if the environmental space is more limited or there is an object of interest, such as a table or poster, people are likely to position themselves differently. This would almost certainly have a negative effect on the prediction of our classifiers. An extension would be to collect further data in such conditions and extend the approach here to factor in space limitations and/or objects as additional inputs. Other potentially relevant aspects of conversational group context were also missing from the datasets we studied. For example, our classifiers were not trained for situations where an extra person may stand within the personal space of another participant (e.g., a child with their parent). Finally, we note that our experimental validation was limited in using only 8 participants. Currently, the COVID-19 pandemic is introducing extreme challenges towards studying in-person group interactions with robots and we hope to conduct further in-person experiments in the future.

X. Conclusion

We presented a new approach for understanding F-formations by identifying "regions"—probabilistic areas where people are likely to stand in a group of a given size—relative to a single "anchor" agent (person or robot). We developed a system comprised of two classifiers using this insight. The first predicts the total number of people in a F-formation based on the positions of only two people in the group. In the case of an undetected person, the second classifier predicts that person's position. Our results demonstrate the potential of our approach in improving how social robots may detect and reason about F-formations.

Acknowledgment

We thank James Kennedy for his collaboration on earlier aspects of this work and the ATLAS Institute at the University of Colorado Boulder for their support during our laboratory validation.

REFERENCES

[1] L. Acosta, E. González, J. N. Rodríguez, A. F. Hamilton *et al.*, "Design and implementation of a service robot for a restaurant," *International Journal of Robotics & Automation*, vol. 21, no. 4, p. 273, 2006.

[2] C. Datta, A. Kapuria, and R. Vijay, "A pilot study to understand requirements of a shopping mall robot," in *Proceedings of the 6th international conference on Human-robot interaction*. ACM, 2011, pp. 127–128.

[3] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita, "An affective guide robot in a shopping mall," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. ACM, 2009, pp. 173–180.

[4] H. Osawa, A. Ema, H. Hattori, N. Akiya, N. Kanzaki, A. Kubo, T. Koyama, and R. Ichise, "What is real risk and benefit on work with robots?: From the analysis of a robot hotel," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2017, pp. 241–242.

[5] E. Zalama, J. G. García-Bermejo, S. Marcos, S. Domínguez, R. Feliz, R. Pinillos, and J. López, "Sacarino, a service robot in a hotel environment," in *ROBOT2013: First Iberian Robotics Conference*. Springer, 2014, pp. 3–14.

[6] H. S. Ahn, S. Zhang, M. H. Lee, J. Y. Lim, and B. A. MacDonald, "Robotic healthcare service system to serve multiple patients with multiple robots," in *International Conference on Social Robotics*. Springer, 2018, pp. 493–502.

[7] D. Bohus, C. W. Saw, and E. Horvitz, "Directions robot: in-the-wild experiences and lessons learned," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 637–644.

[8] J. Rios-Martinez, A. Spalanzani, and C. Laugier, "From proxemics theory to socially-aware navigation: A survey," *International Journal of Social Robotics*, vol. 7, no. 2, pp. 137–153, 2015.

[9] F. Schneemann and P. Heinemann, "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, 2016, pp. 2243–2248.

[10] S. Nasihati Gilani, D. Traum, A. Merla, E. Hee, Z. Walker, B. Manini, G. Gallagher, and L.-A. Petitto, "Multimodal dialogue management for multiparty interaction with infants," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 5–13.

[11] M. Vázquez, E. J. Carter, B. McDorman, J. Forlizzi, A. Steinfeld, and S. E. Hudson, "Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2017, pp. 42–52.

[12] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani, "Multi-scale f-formation discovery for group detection," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 3547–3551.

[13] F. Setti, C. Russell, C. Bassetti, and M. Cristani, "F-formation detection: Individuating free-standing conversational groups in images," *PloS one*, vol. 10, no. 5, p. e0123783, 2015.

[14] E. T. Hall, *The hidden dimension*. Garden City, NY: Doubleday, 1966, vol. 609.

[15] A. Kendon, *Conducting interaction: Patterns of behavior in focused encounters*. CUP Archive, 1990, vol. 7.

[16] L. Tong, A. Serna, S. Pageaud, S. George, and A. Tabard, "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*, 2016, pp. 318–329.

[17] P. Marshall, Y. Rogers, and N. Pantidi, "Using f-formations to analyse spatial patterns of interaction in physical environments," in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 2011, pp. 445–454.

[18] E. De Stefani and L. Mondada, "Reorganizing mobile formations: When "guided" participants initiate reorientations in guided tours," *Space and Culture*, vol. 17, no. 2, pp. 157–175, 2014.

[19] T. Ballendat, N. Marquardt, and S. Greenberg, "Proxemic interaction: designing for a proximity and orientation-aware environment," in *ACM International Conference on Interactive Tabletops and Surfaces*, 2010, pp. 121–130.

[20] N. Marquardt, K. Hinckley, and S. Greenberg, "Cross-device interaction via micro-mobility and f-formations," in *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 2012, pp. 13–22.

[21] H. Hung and B. Kröse, "Detecting f-formations as dominant sets," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 231–238.

[22] M. Swofford, J. Peruzzi, N. Tsoi, S. Thompson, R. Martín-Martín, S. Savarese, and M. Vázquez, "Improving social awareness through dante: Deep affinity network for clustering conversational interactants," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW1, pp. 1–23, 2020.

[23] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[24] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7297–7306.

[25] T. Yu, S. Lim, K. Patwardhan, and N. Krahnstoever, "Monitoring, recognizing and discovering social networks," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1462–1469.

[26] M. J. V. Leach, R. Baxter, N. M. Robertson, and E. P. Sparks, "Detecting social groups in crowded surveillance videos using visual attention," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2014.

[27] W. Choi, Y. W. Chao, C. Pantofaru, and S. Savarese, "Discovering groups of people in images," in *ECCV*, 2014.

[28] O. Brdiczka, J. Maisonnasse, and P. Reignier, "Automatic detection of interaction groups," in *Proceedings of the 7th international conference on Multimodal interfaces*, 2005, pp. 32–36.

[29] T. Choudhury and A. Pentland, "The sociometer: A wearable device for understanding human networks," in *CSCW'02 Workshop: Ad hoc Communications and Collaboration in Ubiquitous Computing Environments*, 2002.

[30] N. Eagle and A. S. Pentland, "Reality mining: sensing complex social systems," *Personal and ubiquitous computing*, vol. 10, no. 4, pp. 255–268, 2006.

[31] T. Gan, Y. Wong, D. Zhang, and M. S. Kankanhalli, "Temporal encoded f-formation system for social interaction detection," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 937–946.

[32] G. Groh, A. Lehmann, J. Reimers, M. R. Frieß, and L. Schwarz, "Detecting social situations from interaction geometry," in *2010 IEEE Second International Conference on Social Computing*. IEEE, 2010, pp. 1–8.

[33] L. Feng and B. Bhanu, "Understanding dynamic social grouping behaviors of pedestrians," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 2, pp. 317–329, 2014.

[34] T. Linder and K. O. Arras, "Multi-model hypothesis tracking of groups of people in rgb-d data," in *17th International Conference on Information Fusion (FUSION)*. IEEE, 2014, pp. 1–7.

[35] A. Matic, V. Osmani, and O. Mayora-Ibarra, "Analysis of social interactions through mobile phones," *Mobile Networks and Applications*, vol. 17, no. 6, pp. 808–819, 2012.

[36] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino, "Social interactions by visual focus of attention in a three-dimensional environment," *Expert Systems*, vol. 30, no. 2, pp. 115–127, 2013.

[37] D. O. Olguín, B. N. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland, "Sensible organizations: Technology and methodology for automatically measuring organizational behavior," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 43–55, 2008.

[38] D. Wyatt, T. Choudhury, and J. Bilmes, "Conversation detection and speaker segmentation in privacy-sensitive situated speech data," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[39] M. Luber and K. O. Arras, "Multi-hypothesis social grouping and tracking for mobile robots." in *Robotics: Science and Systems*, 2013.

[40] M. Vázquez, "Reasoning about spatial patterns of human behavior during group conversations with robots," Ph.D. dissertation, Carnegie Mellon University, 2015.

[41] B. Scassellati, J. Brawer, K. Tsui, S. Nasihati Gilani, M. Malzkuhn, B. Manini, A. Stone, G. Kartheiser, A. Merla, A. Shapiro *et al.*, "Teaching language to deaf infants with a robot and a virtual human," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[42] H. Hung, G. Englebienne, and L. Cabrera Quiros, "Detecting conversing groups with a single worn accelerometer," in *Proceedings of the 16th international conference on multimodal interaction*, 2014, pp. 84–91.

[43] S. Alletto, G. Serra, S. Calderara, F. Solera, and R. Cucchiara, "From ego to nos-vision: Detecting social relationships in first-person views," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 580–585.

[44] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1226–1233.

[45] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino, "Social interaction discovery by statistical analysis of f-formations." in *BMVC*, vol. 2, 2011, p. 4.

[46] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino, "Detecting conversational groups in images and sequences: A robust game-theoretic approach," *Computer Vision and Image Understanding*, vol. 143, pp. 11–24, 2016.

[47] N. Sanghvi, R. Yonetani, and K. Kitani, "Learning group communication from demonstration," in *Workshop on Models and Representations for Natural Human-Robot Communication at the*, 2018.

[48] D. Bohus and E. Horvitz, "Models for multiparty engagement in open-world dialog," in *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, ser. SIGDIAL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 225–234. [Online]. Available: http://dl.acm.org/citation.cfm?id=1708376.1708409

[49] What's an average shoulder width? [Online]. Available: https://www.healthline.com/health/average-shoulder-width#average-shoulder-width

[50] S. Thompson, A. Gupta, A. W. Gupta, A. Chen, and M. Vázquez, "Conversational group detection with graph neural networks," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 248–252.

[51] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung, "The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates," *IEEE Transactions on Affective Computing*, 2018.

[52] G. Zen, B. Lepri, E. Ricci, and O. Lanz, "Space speaks: towards socially and personality aware visual surveillance," in *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*, 2010, pp. 37–42.

[53] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews *et al.*, "Panoptic studio: A massively multiview system for social interaction capture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 190–204, 2017.

[54] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe, "Salsa: A novel dataset for multimodal group behavior analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1707–1720, Aug 2016.

[55] "Reform: Recognizing f-formations for social robots," https://arxiv.org/pdf/2008.07668.pdf, 2020, accessed: 2020-10-01.

[56] "The Resistance (game)," https://en.wikipedia.org/wiki/The_Resistance_(game), accessed: 2019-05-03.