Recognizing F-Formations in the Open World

Hooman Hedayati*
University of Colorado Boulder
hooman.hedayati@colorado.edu

Daniel Szafir
University of Colorado Boulder
daniel.szafir@colorado.edu

Sean Andrist
Microsoft Research, Redmond
sandrist@microsoft.com

***Abstract*—A key skill for social robots in the wild will be to understand the structure and dynamics of conversational groups in order to fluidly participate in them. Social scientists have long studied the rich complexity underlying such focused encounters, or *F-formations*. However, current state-of-the-art algorithms that robots might use to recognize F-formations are highly heuristic and quite brittle. In this report, we explore a data-driven approach to detect F-formations from sets of tracked human positions and orientations, trained and evaluated on two openly available human-only datasets and a small human-robot dataset that we collected. We also discuss the potential for further computational characterization of F-formations beyond simply detecting their occurrence.**

I. Introduction

In order for social robots to fluidly participate in spontaneous multi-party conversational interactions that may arise in the open world environment in which they are deployed, they will need to learn how such conversational groups are formed, how they are shaped, and how they evolve. Such encounters were operationalized by Kendon as *F-formations* and social scientists have long explored the intricate nuances of their structure and dynamics [2]. For a robot to comfortably take part in such groups, it must first be able to detect when they occur among the humans in its environment and with itself.

Previous approaches to F-formation detection are generally heuristic, optimization-based algorithms with hand-tuned parameters [1], [4]. Researchers have shown great promise in applying such techniques to human-computer and human-robot interaction, often by jointly reasoning about human head and body orientations [3], [5]. However, these algorithms are difficult to generalize and tune across different environments and population mixtures of humans and robots. In this report, we seek to address this limitation by exploring an array of machine learning classification models on an openly available dataset of human-human labeled F-formations [1]. We also apply these models to a small amount of acted human-robot data collected from a deployed system in order to assess feasibility (Figure 1). Finally, we propose ways to computationally characterize F-formations as a means to make comparisons across datasets, which we believe will be important for robots to act appropriately in different social and physical contexts.

II. Approach

One of the challenges towards using machine learning models for detecting F-formations is the lack of labeled data to train models. Moreover, the problem of detecting F-formations directly is not conducive to standard classification techniques due to their variable size. To address both these issues, this report suggests inspecting F-formations at the pairwise level by deconstructing the annotated frames into all possible pairs of people. Given a scene of n people with measured positions and orientations, our algorithm involves the following steps:

Dataset Deconstruction: In the first step, the algorithm deconstructs the n individuals' data and create $n(n-1)/2$ pairwise data points. For every two people in the scene, the algorithm creates a data point with two new features: *Distance* and *Effort Angle*. Distance is defined as Euclidean distance between two individuals, which is an intuitive feature because people tend to have a conversation with others at a comfortable distance. Effort Angle defines how much body rotation would be required for the two people to both be pointed directly at each other. The range is between 0 and 2π: 0 means the two people are already facing each other, while 2π means they are facing in opposite directions from each other.

Pairwise Classification: The pairwise data are then used to train a binary classifier with labels indicating whether the two people are inside or outside of an f-formation in that frame (labels of 1 or -1). This binary classification model can then be applied to new data that has also been deconstructed into pairs using the previous step. Model predictions on the pairwise data are used to create a *Relation Matrix* M_R. With n people in the scene, M_R is an $n \times n$ symmetric matrix in which for each two people P_i and P_j, a_{ij} and a_{ji} are equal to the predicted label and the diagonal entries are all 1.

Reconstruction: The final step is to reconstruct the pairwise data into full F-formation sets. Because the classifier will not be perfect, there might be inconsistencies across the pairwise predictions. For example, in a frame with three individuals P1, P2, and P3, the pairwise classifications might indicate that P1-P2 and P2-P3 are in F-formations, but not P1-P3. To resolve this problem, we developed a voting algorithm. The high-level idea is that we accumulate evidence into a larger F-formation when it involves believing more pairwise evidence than would be disbelieved were the formations left separated.

Let row i of M_R indicate P_i's belief about their F-formation status with all others in the scene, which we refer to as B_i. The voting scheme algorithm finds B_i and B_j which have the maximum number of elements in their intersection ($B_i \cap B_j$). Then $B_i \cup B_j$ will be the first F-formation detected by the algorithm, and B_i and B_j are deleted from the rows and columns associated with P_i and P_j. This process repeats till

*This work was conducted as part of an internship at Microsoft Research.

there is no B left in the M_R. As an example, if B_1=P_1,P_2,P_3, B_2 = P_1,P_2,P_3,P_4, and B_3=P_1,P_2,P_3, it is more likely that B_3 is incorrect rather than both B_1 and B_2, *i.e.*, it is more likely that the classifier made one mistake rather than two.

III. Evaluation

To evaluate our proposed approach, we first trained a set of ML classification models on the SALSA dataset [1]. We randomly divided into train and test sets of 80% and 20% respectively. SALSA was chosen from the set of currently openly available datasets because it has a relatively large amount of annotated frames and includes large scenes of people (18 per frame). We report the pairwise training accuracy with 5-fold cross-validation for three ML models: Weighted KNN, Bagged Trees, and Logistic Regression (Table I).

After reconstructing the pairwise results into F-formations on both the train and test sets, we compared to a majority class baseline and our implementation of the existing state-of-the-art algorithm, Graph-Cuts [4]. We used parameters of $mdl = 30000$ and $stride = 0.7$ for the latter as provided in the open-source code. Precision, recall, and F1 scores for the resulting F-formation results are shown in Table I with parameter $T = 2/3$ as described by Setti et al. [4] (F-formations with 2/3 match to ground truth are counted as correct). All three models perform better than Graph-Cuts on this dataset, although parameter tuning per dataset could potentially increase Graph-Cuts' performance. Bagged trees perform the best on the test set, although logistic regression exhibits the least amount of overfitting.

We also applied our models to a small amount of data collected by acting out a few different spatial configurations on an existing in-the-wild robot system (see Figure 1) and found that they hold promise for recognizing when there is no active F-formation (*e.g.*, people just walking past), F-formations involving only humans (*e.g.*, two people chatting away from the robot), and F-formations involving the robot (*e.g.*, two people intending to interact with the robot).

IV. Discussion & Future Work

This evaluation indicates great promise in accuracy and generalizability for our data-driven approach to detecting F-formations. But beyond simply detecting them as they occur, we also propose that robots will need access to a finer-grained characterization of F-formations in order to fully understand and participate. We introduce $Tightness$ and $Symmetry$ as two such characterizations. $Tightness$ is the average distance between participants and the F-formation's center. $Symmetry$ is the difference in average angular difference between people in the F-formation to what would be a completely symmetric configuration. Large values might indicate that there is an obstacle or situational attractor in the scene, *e.g.* a table or a poster. We are currently exploring how these measures significantly differ across group sizes and in different environments.

Going forward, we first plan to train and test models on larger datasets. We will then integrate these models into a larger pipeline including more sophisticated techniques for tracking human poses in 3D, incorporating uncertainty and active sensing into the reasoning. We will explore a richer set of features such as velocity, head orientation, eye gaze, dialog context, social status, historical features, *etc.*, alongside more sophisticated temporal and online modeling techniques, such as the tightness and symmetry features described above. Through this work, we hope to imbue robots with new methods to reason about complex social dynamics to improve the fluidity and naturalness of human-robot interactions.

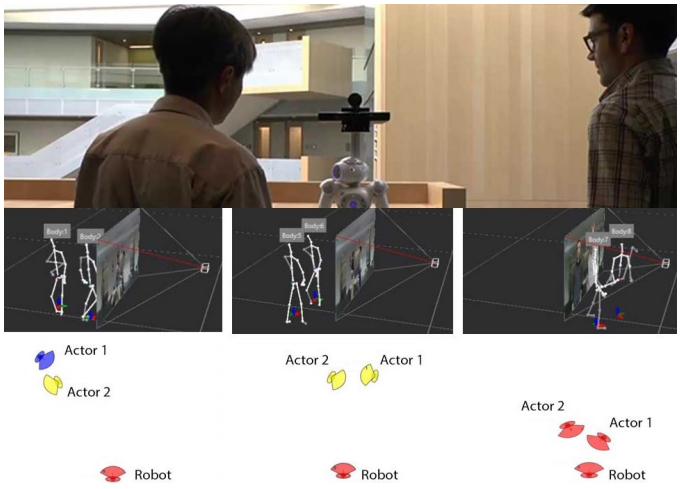

Fig. 1. Top: The robot set up, Middle: 3D visualization of human poses perceived by the robot using Kinect, Bottom: Visualization of classified F-formations indicated by actors sharing the same color.

TABLE I
Results on train/test split of SALSA dataset compared with baseline and an existing state-of-the-art algorithm

	Pairwise Accuracy	**Precision (Train/Test)**	**Recall (Train/Test)**	**F1 (Train/Test)**
Majority Baseline	85.3	100/100	0/0	0/0
Graph-Cuts	N/A	66.2/63.8	64.2/64.2	65.2/64.2
Weighted KNN	92.1	86.5/78.1	99.9/82.7	92.7/80.3
Bagged Tree	**93.3**	**86.3/78.3**	**99.4/84.3**	**92.4/81.2**
Logistic Regression	92.2	73.9/71.3	78.9/78.6	76.3/74.8

References

[1] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, "Analyzing free-standing conversational groups: A multimodal approach," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 5–14.

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

[3] 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.

[4] 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.

[5] 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.