CPSC 290 Final Report
Improving Social Awareness and Group Detection through Deep Learning

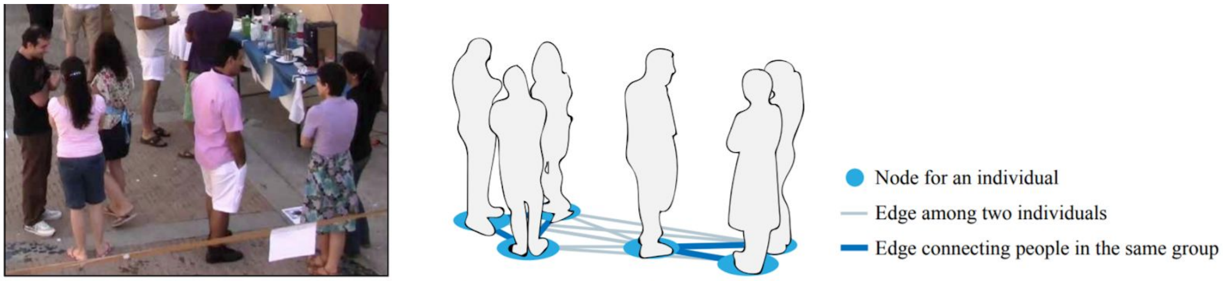
Author: Abhijit Gupta, Yale College '23
Advisor: Dr. Marynel Vazquez, Department of Computer Science

Abstract

Automatic group detection using computer vision is critical to surveillance systems, socially-aware mobile systems, interactive displays, and more. In the field of human-robot interaction, group detection is necessary for optimal verbal and non-verbal behavior. Explicitly modeled heuristics are often used to group people by spatial arrangement, but these methods do not generalize well to more complex scenarios. With increasingly available high-quality data inputs, powerful deep learning methods can improve the detection of conversational groups in comparison to prior approaches. The current state-of-the-art technique uses a Deep Affinity Network (DANTE) and the Dominant Set (DS) clustering algorithm to create group predictions. The previous implementation of the DANTE algorithm was designed to train on a pre-compiled dataset with six people in each frame and a limited input space. In preparation for training the model on live feed from the Shutter Robot, with variable people and an expanded feature space, I re-implemented the DANTE algorithm with an emphasis on scalability and robustness. This new implementation can handle variable people in each frame and scales easily as new input features are incorporated. Using previously collected body pose information for the Cocktail Party dataset, I trained a model on the expanded dataset as a proof of concept. The expanded dataset improved T=1 F1 scores from 0.58 to 0.69 and T=⅔ F1 scores from 0.80 to 0.86. Future directions include collecting and training on a new dataset using the Shutter Robot and fine tuning model structure and hyperparameters for increased performance.

Background

Automatic group detection using computer vision is critical to surveillance systems, socially-aware mobile systems, interactive displays, and more. In the field of human-robot interaction, group detection is necessary for optimal verbal and non-verbal behavior [1]. Group conversation can be visually recognized by analyzing the location and orientation of all individuals. In this model, each individual is a node on a graph, with edges connecting people in the same group [2]. Figure 1 below gives an example input and output of this process.


● Node for an individual
Edge among two individuals
Edge connecting people in the same group

(a) Scene from the Coffee Break dataset

(b) Interaction Graph

Figure 1: Sample Input and Output [2]

Currently, explicitly modeled heuristics are often used to group people by spatial arrangement, e.g., [3]. But these methods do not generalize well to more complex scenarios due to the rigid assumptions made. With increasingly available high-quality data inputs, powerful deep learning methods can improve the detection of conversational groups in comparison to prior approaches. The current state-of-the-art technique uses a Deep Affinity Network (DANTE) and the Dominant Set (DS) clustering algorithm to create group clustering predictions [2]. The data analysis pipeline is shown in Figure 2 below.

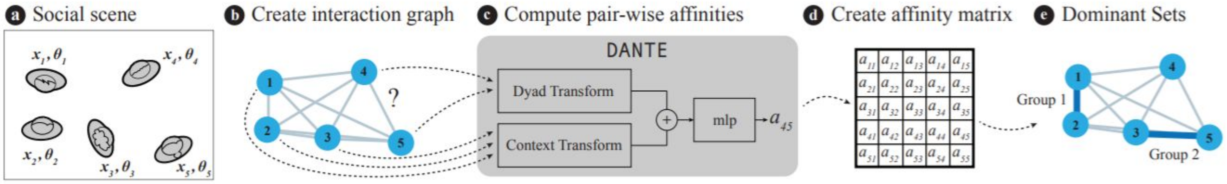

a Social scene
x_1, θ_1
x_4, θ_4
x_2, θ_2
x_3, θ_3
x_5, θ_5
b Create interaction graph
1
4
?
2
3
5
c Compute pair-wise affinities
DANTE
Dyad Transform
Context Transform
+
mlp
a_{45}
d Create affinity matrix
a_{11} a_{12} a_{13} a_{14} a_{15}
a_{21} a_{22} a_{23} a_{24} a_{25}
a_{31} a_{32} a_{33} a_{34} a_{35}
a_{41} a_{42} a_{43} a_{44} a_{45}
a_{51} a_{52} a_{53} a_{54} a_{55}
e Dominant Sets
Group 1
Group 2
1
4
2
3
5

Figure 2: DANTE Pipeline [2]

In Figure 2 (a), the scene is captured and spatial feature vectors for each individual are extracted. An interaction graph is created in (b) and DANTE uses both a Dyad and a Context Transform with several multilayer perceptrons (c) to compute weights that represent the likelihood that two individuals are part of the same conversational group based on their spatial features. These weights are then organized into an affinity matrix, showin in (d). Finally, Dominant Sets (e) is applied to the matrix to determine the predicted group structure. The DANTE algorithm is preferred for its deep learning components and ability to contextualize both individual interactions and the overall social scene. Compared to previous benchmarks tested on publicly available datasets, DANTE resulted in greater precision and recall with statistical significance [2]. We build upon the existing DANTE implementation in this work.

Algorithms such as DANTE are evaluated on several publicly available datasets. Each source consists of a video recording, and for each frame, extracted features and ground truth group annotations. The Cocktail Party Dataset [4], SALSA Dataset [5], and Coffee Break Dataset [6] are three commonly used references. An example input from the Cocktail Party Dataset is shown below. The raw picture is transformed into a grid representation with each participant's location and orientation. While these datasets represent a suitable starting point, each is limited by a combination of small sample sizes, limited feature spaces, and noisy data. An additional goal of this work was to develop and share a larger dataset with an expanded feature space.

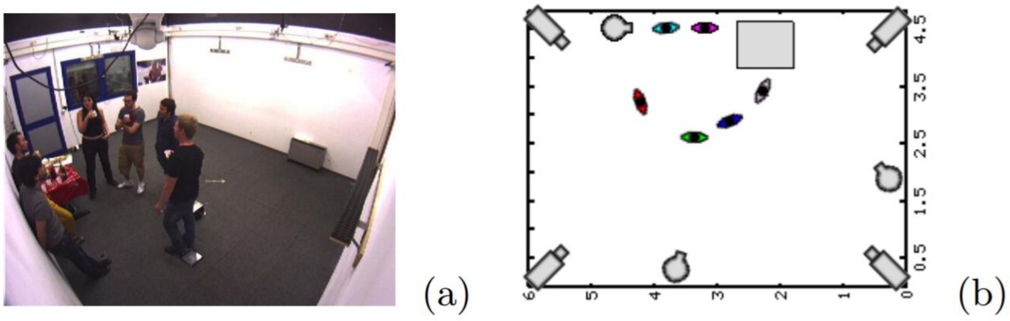


Figure 3: Cocktail Party Example Frame [4]

Project and Results

The previous implementation of the DANTE algorithm was designed to train on a pre-compiled dataset with six people in each frame and a limited input space. In preparation for training the model on live feed from the Shutter Robot, with variable people and an expanded feature space, I re-implemented the DANTE algorithm with an emphasis on scalability and robustness. This new implementation can handle variable people in each frame and scales easily as new input features are incorporated. Due to the coronavirus pandemic, I was unable to implement the model in ROS to use on the Shutter Robot. However, using previously collected body pose information for the Cocktail Party dataset, I trained a model on the expanded dataset as a proof of concept.

For a given frame, the DANTE algorithm outputted a predicted set of groups, which was then compared against the ground truth annotations. For a threshold T, a frame was deemed correct if (correct groups / total groups) >= T. Precision, recall, and f1-scores were compiled with T=⅔ and T=1. As the Cocktail Party dataset contains only 320 frames, five-fold cross validation was used, with metrics calculated on each fold and averaged together for the final results.

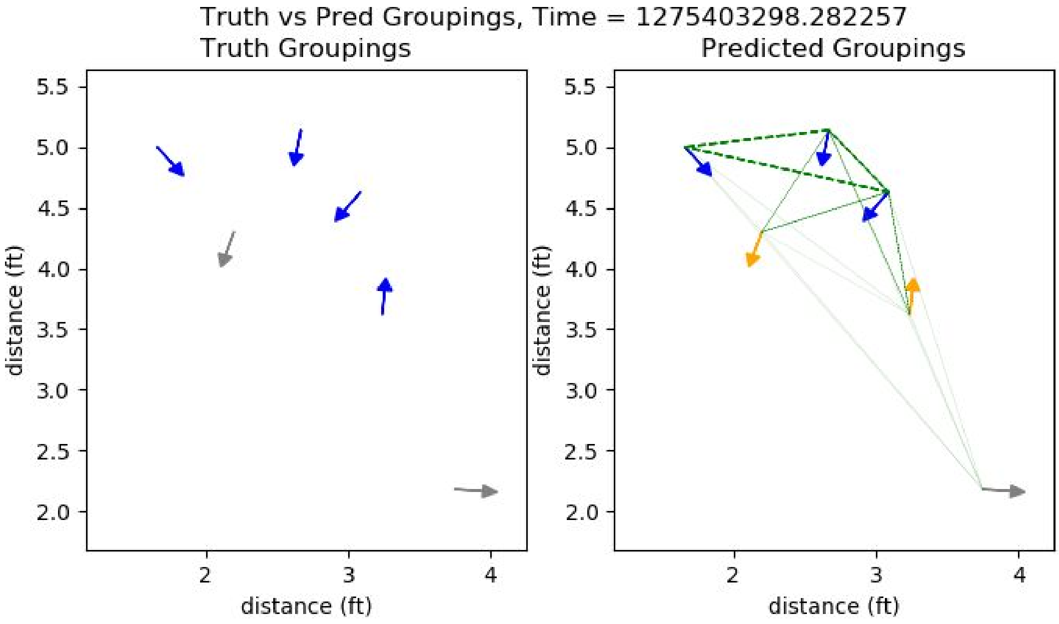


Figure 4: Sample Frame Truth Groupings vs Predicted Groupings - colors code groupings, predicted affinities between individuals shown by green line intensity on left

The original dataset consisted of 320 frames with annotations for each person's x-coordinate, y-coordinate, and head orientation. The expanded dataset uses these same frames with additional body orientation annotations. Using the original dataset as a baseline, I compared the T=⅔ F1 and T=1 F1 scores for the DANTE algorithm trained on the expanded dataset. The results are shown below:

Dataset	T=⅔ F1-score	T=1 F1-score
Cocktail Party	0.80	0.58
Expanded Cocktail Party	0.86	0.69

The expanded dataset led to significantly higher classification metrics, especially on the T=1 F1-score, corresponding to exact group predictions. These results show the benefit of a larger input space, and prove the new implementation can incorporate these additional variables.

Future Directions

Having validated the new DANTE implementation with the expanded dataset test, the next step would be to use the Shutter Robot to collect a new dataset. With variable input size, an expanded feature space using skeletal pose from the Kinect Azure cameras, and a larger sample size (more annotated frames), we should be able to significantly improve the classification accuracy of the DANTE algorithm. This dataset would also be shared to advance face-formation research.

Even with the current dataset, there are still avenues to improve model performance. Comparing against the current model as a baseline, I can experiment with the model structure and tune hyperparameters. I am also looking into face-formation heuristics that can be implemented alongside the DANTE model. Lastly, having predicted current groupings, a new model can be developed to position a robot to engage in an ongoing conversation. These advancements will allow autonomous social agents to better reason about and interact with their environments.

References

[1] 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

[2] 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

[3] Francesco Setti, Chris Russell, Chiara Bassetti, and Marco Cristani. 2015. F-formation detection: Individuating freestanding conversational groups in images. PloS one 10, 5 (2015), e0123783

[4] Gloria Zen, Bruno Lepri, Elisa Ricci, and Oswald Lanz. 2010. Space speaks: towards socially and personality aware visual surveillance. In Proceedings of the 2010 ACM International Workshop on Multimodal Pervasive Video Analysis (MPVA). ACM, 37–42

[5] 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 Transactions on Pattern Analysis and Machine Intelligence (PAMI) 38, 8 (Aug 2016), 1707–1720. https://doi.org/10.1109/TPAMI.2015.2496269

[6] 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 Proceedings of the 2011 British Machine Vision Conference (BMVC). BMVA Press, 23.1–23.12.