Autonomous Group Detection, Delineation, and Selection for Human-Agent Interaction*

Ben Wright[1], J. Malcolm McCurry[2], Wallace Lawson[3], and J. Gregory Trafton[3]

[1] NRC Postdoctoral Associate, US Naval Research Laboratory
benjamin.wright.ctr@nrl.navy.mil
[2] Peraton, Washington D.C., USA jmccurry@peraton.com
[3] Navy Center for Applied Research in Artificial Intelligence, US Naval Research Laboratory
{ed.lawson,greg.trafton}@nrl.navy.mil

Abstract. If a human and a robot team need to approach a specific group to make an announcement or delivery, how will the human describe which group to approach, and how will the robot approach the group? The robots will need to take a relatively arbitrary description of a group, identify that group from onboard sensors, and accurately approach the correct group. This task requires the robot to reason over and delineate individuals and groups from other individuals and groups. We ran a study on how people describe groups for delineation and identified the features most likely used by a person. We then present a framework that allows for an agent to detect, delineate, and select a given social group from the context of a description. We also present a group detection algorithm that works on a mobile platform in real-time and provide a formalization for a Group Selection Problem.

Keywords: Group Description · Group Identification · Qualitative Reasoning · Proxemics

1 Introduction & Motivation

Individual and group detection have long been an area of research [14, 9, 22]. In many cases, these detections are maintained internal to the agent or given as annotations to video or images for a human teammate to use. Group detection becomes more important when we think about it in terms of how humans describe groups to each other. Usually, it is not the case that we have a label for the group to describe to a friend or teammate. Normally, groups are described (e.g. "The group by the window." or "The large group right next to us."). Through these types of descriptions, we see that grouping is contextual [15, 13]. Therefore, if we wish to have agents cooperating in Human-Agent teams, we should ensure that agents can handle contextual group descriptions.

* This research was performed while BW held an NRC Research Associateship award at NRL. This research was funded by ONR and OSD to GT.

The contributions of this paper are (1) we show that people use group descriptions when labels are not present, (2) we introduce a very simple group detection algorithm, (3) we establish a Group Selection Problem, and (4) we provide a basic implementation solution and evaluation to both group detection and selection.

2 Background

This work touches on a number of different fields of research, we provide some background from proxemics in terms of representation, recognition, and detection efforts.

2.1 Proxemics Representations of Groups

First discussed in [6], Proxemics is the study of spatial interaction among humans. Proxemics has had a growing interest in regards to research for Human-Robot Interaction (HRI) [23, 17, 21, 4, 3, 12]. Currently, the study of proxemics in this context falls into two camps: F-Formations or not F-Formations.

Face-to-face formations (F-formations) have received a lot of attention in terms of group proxemic research [23, 17, 27]. Originally discussed in [8], F-Formations describe different spatial arrangements of people given the number of people interacting. [9] defines it as,"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". F-formations focus on three different spatial zones in relation to a group: o-space, p-space, and r-space. O-space exists inside the perimeter of the group, p-space is the perimeter of the group space and r-space being the space outside of p-space.

Aside from F-Formations, a number of other methods are being used to study and represent groups. [21] uses optical flow for active egocentric group detection in motion. [4] built a representation using qualitative spatial descriptors on top of F-formations. This is done by using logical constraint rules to determine the interactions of two-pair F-formations to create larger group formations. This is then used as a way to reason how a potential robot can join the group in a new formation pattern. [3] uses Qualitative Trajectory Calculus (QTC) to encode interactions in an HRI setting. This allows for qualitative reasoning on movement and is used in environments with trajectory detection and planning [12]. Using individual person detection and tracking, [12] builds up social network graphs between all the individuals and uses various pruning methods to then detect proper groups.

2.2 Group Recognition & Detection Work

Previous work in group detection has generally followed two different constraints: stationary vs moving groups. For stationary groups, a number of algorithms use video stills. For instance, [27] uses stills to detect F-formations based on

dominant sets [23, 7]. This is where individuals in the image are converted to a graph and given proximity weights, these weights are then used to detect the groups. Tracking over multiple images or video has had a wide variety of use in group detection. Having multiple images allows for a wide variety of trajectory and velocity tracking. [21], which follows the work of [12], has a discussion about maintaining identities as groups and agents move around the scene. Other video or moving group detections can be seen in [10, 19, 20].

3 Group Description Experiment

Our goal in this experiment was to explore how people identify and differentiate groups within a space that contains multiple groups. There were theoretical reasons from both the cognitive sciences and the interaction sciences that this experiment will help us answer. From the cognitive sciences, there are no theories or expectations about how people will deal with identifying a group of diverse individuals and whether they will over-describe a group. From the HRI perspective, there is no clear understanding of what features people will use to describe a group; knowing the features that are used allows us as roboticists to tailor our sensors and perceptual systems to what features are going to be the most useful. For example, if people commonly identify most group features and several individual features, most perceptual systems should be adequate. In contrast, if people primarily use a specific single feature, it would behoove us to make sure that our robot perceptual systems can deal with that feature.

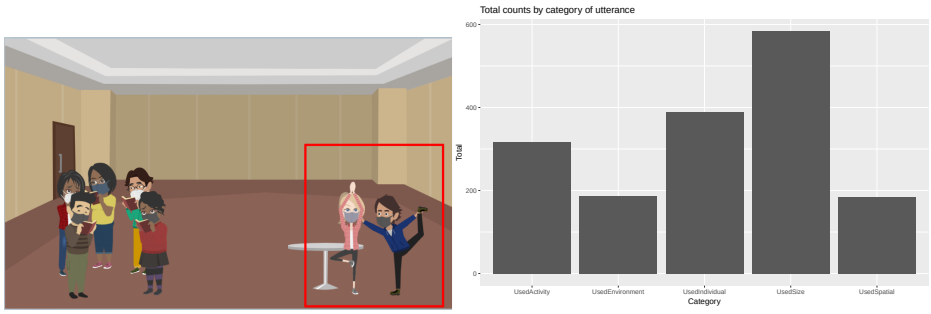
3.1 Setup

In this experiment, we were interested in three distinct questions concerning how people identify groups. First, because groups have both group features and individuals within a group that can be identified, we wanted to document whether people used group or individual features more. Second, we wanted to identify how Gricean people are in their group descriptions. A purely Gricean approach would mean identifying a single unique feature that differentiated the groups and using that feature: no more and no less. Finally, we were interested in determining what features people used to identify groups, and whether those features were random or systematic.

97 participants from Amazon Mechanical Turk were paid $2.00 to answer questions about how to identify a group. They were told they were working with a robot to deliver snacks to individuals. Participants were randomly assigned to either the Same Perspective condition where they were told that the robot could either see from the same perspective that they could (n=53) or the Different Perspective condition where they were told the robot would be coming from an unknown direction so could not use words like “left” or “right” (n=44).

Images were constructed using Vyond[4] to contain 2 distinct groups with differing group features. Group features that were explicitly manipulated were size

[4] `https://www.vyond.com/`



(a) An example image participants were shown.

(b) Feature breakdown by Utterance, from the left: Activity, Object, Individual, Size, Proximity

Fig. 1: Experiment Image and Results

(same or different with sizes of 2 or 5), object (close or far from a group), activity (same or different with dancing, stretching, talking, or reading), or proximity (close or far from each other). The selected group had an obvious rectangle drawn around it and could be on either the left or the right (counterbalanced). An example of an image is shown in Figure 1a. Each participant saw 10 images. Because it was impossible to ask each participant to see one of each variable, we gave each participant two examples of 0-4 differences from manipulated features (size, object, activity or proximity), using a Latin-square design to keep the presence of features as equal as possible. This approach allowed us to use regression analyses to extract patterns from the participants. After collecting demographic information, each participant was given a brief description of the task and then asked to type in an English description of where to tell the robot to go. Each image also reminded the participant that the robot could see the same scene as the participant or that the robot would be approaching from an unknown location. After finishing the experiment, participants were debriefed.

3.2 Results

We used a combination of keyword extraction and hand-coding to identify which feature(s) a participant used to identify each group. Features coded were size ("The group of 5"), object ("The group closest to the table"), activity ("The dancing group"), spatial ("The group to the left"), or individual ("The group with the blond woman"). Multiple features of different types were each coded, though multiple features of the same feature were only coded once (e.g., "The group on the left with the tall blond woman" would be coded as spatial and individual, even though there are two references to individual traits).

In general, people used more group features (size, spatial, object, activity) than individual features when describing a group, $\chi^2(1) = 503.6$, $p < 0.001$

(829 group vs 133 individual).[5] To explore how Gricean participants were, we examined the raw number of features that could be used to differentiate groups; if people used a single distinguishing feature to differentiate groups, it was coded as Gricean. If people used more than was needed, it violated the Gricean maxim of quantity. There were very few instances of people not using a description that did not differentiate groups ($< 1\%$). We found that people were very likely to over-describe groups (66% vs. 34%), $\chi^2(1) = 84.7$, $p < 0.001$. People strongly violated Gricean maxim of quantity in this experiment.

To examine whether people were random in their choice of features to describe and differentiate groups or whether they had systematic preferences, we examined which features were used, collapsing across other variables because they showed no significant effect. As Figure 1b suggests, people did not choose features at random to differentiate groups, $\chi^2(4) = 329.6$, $p < 0.001$. A post-hoc Tukey test showed a particular order that people preferred: Size > Individual > Activity > Object = Proxemic.

As hoped, this experiment helped answer several theoretical and applied questions from cognitive science and HRI. First, people seem to prefer to use group features rather than features of individuals to differentiate groups, though both are used frequently. Second, this experiment provides some expectations about how people will address a robot to approach a group. Specifically, people do not use a random set of features to differentiate groups, instead choosing to have a preferred order: some features are much more likely to occur than others. Finally, people are quite non-Gricean when describing a group, frequently providing more information than needed to differentiate one group from another.

4 Group Detection & Representation

With these experimental results in mind, we can begin to work on group detection. From the results above, we know that group size is *very important* to keep accurate. From previous works in group recognition/detection there are two basic ideas, using pose information or not using pose information. However, pose is captured in those datasets within a "smart-room" environment with multiple cameras at useful angles (e.g.,[17]). Unfortunately, these assumptions do not work for our target domain, mobile robotics. For a single robot, the sensors and computing power available is usually far less than available in a typical smart-room – a laptop and a standard CCD.

In contrast, it is far easier to identify the location of a person (regardless of their pose) in most scenes, even when they are partially occluded or are not facing the camera. To show this, we ran OpenPose [1] (a common method to extract pose in real-time) on a series of images. We also ran YOLO [14] to extract where people were in the scenes. In 100% of the cases, YOLO identified more people than OpenPose. For example, in one crowded image environment YOLO detected all but 1 of 12 people while OpenPose failed to give pose information

[5] A linear effects mixed model shows a similar result while taking into account the multiple random effects. A later report will provide a fuller description.

to 4 of the 12. Missing the Pose Information of 33% is significant when it comes to the tightly coupled nature of pose-based detection methods.

Therefore, our designs are not focused on utilizing pose information as maintaining an accurate number of individuals is more important than maintaining an accurate number of poses.

4.1 Agent Detection

To detect people, we use the deep convolutional neural network YOLO [14]. YOLO predicts both the location and the classification of known objects in the image. It does this by first subdividing the image into grid cells, then predicts the most likely size and location for an object in each cell. This network was trained on the MS-COCO dataset [11], which includes a variety of classes. Additionally, we could utilize a robot's on-board range/bearing detection to determine placement of a detected person. From these range and bearing values, we can project the image detection onto a grid.To get group definitions, we assume we are given accurate representations of the agents and their locations within an image - though not necessarily the agent's bearing or facing.

4.2 Definition of a Group

[26] defines a group from video footage as a group of individuals within a certain space threshold and maintaining that threshold for a specific time threshold. [21, 12] defines groups as "two or more people in close proximity to one another with a common motion goal." F-formations utilizes spatial and directional information to determine groups based on o- and p-space[17, 24, 4]. Additionally, f-formations focus on the context of *conversational groups*. With the idea that conversational "space" defines the group by individuals "facing" in a specific way based on certain formation types. Finally, some research leaves the definition of group undefined and utilizes clustering techniques to have groups detected. This research does focus on trajectories heavily though [19, 18].

From these various discussions of groups, we can pull out a few recurring concepts: spatial-temporal relations, motion/direction relations, and shared goals. Not all images can easily be broken down into direction/facing information for all individuals, likewise temporal relations cannot be done from still frames. Furthermore, goals and intent are a very hard problem that is its own research area. Therefore, we try to focus our definition of groups to be purely spatial, but limit it to stationary groups. A *stationary group* is the largest number of agents (greater than or equal to 2) in an environment that are a given distance apart from each other that does not break the standard deviation of the average distance between all agents in the stationary group. It is also assumed that an agent is only a member of one group at a time.

4.3 Group Detection for Stationary Groups

We devised a *Group Detection Algorithm*, based on this definition. Starting with all agents, we calculate the $distance - pair$ between each Agent. Starting with

Described Thing	In Relation To	Query	Example
the group	size,small	$smallest\_group(Group)$.	The smallest group
the group	size,large	$largest\_group(Group)$.	The largest group
the group	size,specific	$group\_of\_size(Group, N)$.	Group of size 5
the group	with,Person	$group\_with(Group, Agent)$.	Group w/ person in hat

Table 1: Group Description options, with Query Examples

the minimum $distance-pair$, we slowly attempt adding in new agents. If the new $distance-pair$ values are within a threshold of mean $distance-pair$ plus/minus standard deviation, we add the new agent to the overall group. Otherwise, we continue to the next possible agent. When we've run out of possible agents to add, whatever is left we define as a new group. We then remove these agents from the overall list of agents to check and start over with the newest minimum $distance - pair$. We continue doing this until our minimum $distance - pair$ reaches a certain threshold if a minimum threshold is not met then we stop, likewise if we run out of agents for pairs we also stop.

4.4 Group Delineation & Selection

A *group description* is a pairing, (R, P), between a relation, R, and a property, P. A group, G, can satisfy a group description if the property and relation hold true for that group. With these group descriptions, we can now think about how we might differentiate between groups. To delineate groups, we need to know all possible descriptions for a group. A *Group Selection Problem*(GSP) consists of a given Group Description, (R, P), and a domain of agents, groups, and objects. A solution for this would be a group that satisfies the given Group Description in the domain.

5 Implementation & Evaluation

5.1 Implementing the Group Selection Problem

We use SWI-prolog [25] to implement our Group Detection algorithm. This follows along previous work that has used Prolog in HRI and qualitative reasoning [4, 26, 5]. To implement group detection we assume knowledge of individuals. From here, our group detection algorithm is used. Some of the functions not fully defined include $findAllDistances$. This function generates a list of *distance pairs*, (a, b, N), between all agents a, b so we know that the distance between agents a and b is N. This is only done once in the detection algorithm.

Additionally, there are some stop cases for thresholds that are important to know. There are two places where thresholds can change how detection works. The first is when the algorithm only has two agents left, there is a threshold to consider them a group or not. The main loop, that is when there are more than 2 agents, automatically makes a group between two agents. The second

Algorithm	Precision	Recall	F-score
Our Method	0.62	0.21	0.31
Pose-only Baseline from [16]	0.29	0.27	0.28

Table 2: Group Detection Methods in the Cocktail Party Benchmark

threshold has to deal with how the standard deviation works on newly formed groups. When a new group-pair is formed, the standard deviation is 0. This is bad, so we allow for this special case that the standard deviation is related to the distance between those two agents. We ran things between a quarter of the distance to two times the distance between the agents. For later testings, these values were used to fine-tune some of our results.

Following along with our group detection, we also implement the group selection queries in SWI-Prolog. Table 1 lists queries next to each description. These queries can be solved fairly quickly from the properties given. A majority of the group property queries also involve a minimum or maximum value check against certain properties.

In addition to our own scenarios to test group detection and selection, we also use a test-case dataset to compare our group detection algorithm to state-of-the-art algorithms - *Cocktail Party*. The *CocktailParty* dataset, first discussed in [17], is a dataset containing video of a cocktail party in a large room with 7 individuals walking around. There is a main table at one side of the room with drinks and food on it and the rest of the room is fairly open for the people to walk around in. Over the course of 30 minutes, the individuals interact with each other in various groupings.

5.2 Evaluations & Testing

We determined a group to be accurately detected as previous detection strategies in [17, 2]. To test our Group Detection Algorithm, we ran it over a known dataset, *CocktailParty*, and gave its Precision, Recall, and F-Score results in comparison to another pose-free group detection algorithm result in Table 2. Groups in this dataset were annotated manually by an expert every 5 seconds resulting in 320 still frames with group annotations. We tested our algorithm against these 320 frames. When compared to a previous "Pose-only" from [16], we do significantly better with precision and, as a result, have a better F-score as well.

These were all run on a basic laptop running Ubuntu 16.04 with a 2.8GHzx8 i7 Intel processor. The runtime results were taken using the linux *time* command and includes the entire prolog run. Each run was given the agents and objects locations and it ran detection, definition, and selection each time. On average our runs took around $0.035s$. Comparatively, [17] mentions averaging about 15 seconds per frame of video for their detection method and [16] mentions their detection algorithm running in a few milliseconds.

6 Discussion & Conclusion

We provide an alternative approach to group formation that need not rely on head tracking or pose detection which can be difficult in some scenarios. We demonstrate an improvement on current approaches [16, 17] that detect group formation without the use of head pose information.

Future directions with group selection involve on-boarding this to full robotic platforms along with adding selection commands. A final direction to take these group description and selection queries is to use them to clarify and prune imperfect information and confirm certain group descriptions with human counterparts. Additional studies in understanding the priorities and orderings for group descriptions used by humans would be beneficial to ensure that this reversal of queries is more optimally specified.

Acknowledgements

The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. The authors would like to thank Magda Bugajska and Bill Adams in their thoughts on combining the detection to representation on robotic platforms.

References

1. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019)
2. Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., Murino, V.: Social Interaction Discovery by Statistical Analysis of F-Formations. In: Proceedings of the British Machine Vision Conference. pp. 23.1–23.12. BMVA Press (2011), http://dx.doi.org/10.5244/C.25.23
3. Dondrup, C., Bellotto, N., Hanheide, M., Eder, K., Leonards, U.: A Computational Model of Human-Robot Spatial Interactions Based on a Qualitative Trajectory Calculus. Robotics **4**(1), 63–102 (2015)
4. Falomir, Z., Angulo Bahón, C.: A Qualitative Spatial Descriptor of Group-Robot Interactions. In: Leibniz International Proceedings in Informatics. pp. 3–1. Schloss Dagstuhl-Leibniz-Zentrum für Informatik (2017)
5. Falomir, Z., Pich, A., Costa, V.: Spatial Reasoning About Qualitative Shape Compositions. Annals of Mathematics and Artificial Intelligence **88**(5), 589–621 (2020)
6. Hall, E.T.: The Hidden Dimension, vol. 609. Garden City, NY: Doubleday (1966)
7. Hung, H., Kröse, B.: Detecting F-Formations as Dominant Sets. In: Proceedings of the 13th international conference on multimodal interfaces. pp. 231–238 (2011)
8. Kendon, A.: The F-formation System: The Spatial Organization of Social Encounters. Man-Environment Systems **6**(01), 1976 (1976)
9. Kendon, A.: Spatial Organization in Social Encounters: The F-Formation System. Conducting interaction: Patterns of behavior in focused encounters (1990)

10. Khan, S.D., Vizzari, G., Bandini, S., Basalamah, S.: Detection of Social Groups in Pedestrian Crowds Using Computer Vision. In: International Conference on Advanced Concepts for Intelligent Vision Systems. pp. 249–260. Springer (2015)
11. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft Coco: Common Objects in Context. In: European Conference on Computer Vision. pp. 740–755. Springer (2014)
12. Linder, T., Arras, K.O.: Multi-Model Hypothesis Tracking of Groups of People in RGB-D Data. In: 17th International Conference on Information Fusion (FUSION). pp. 1–7. IEEE (2014)
13. Long, M., Rohde, H., Rubio-Fernandez, P.: The Pressure to Communicate Efficiently Continues to Shape Language use Later in Life. Scientific Reports **10**(1), 1–13 (2020)
14. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788 (2016)
15. Rubio-Fernandez, P., Jara-Ettinger, J.: Incrementality and Efficiency Shape Pragmatics Across Languages. Proceedings of the National Academy of Sciences (2020)
16. Sanghvi, N., Yonetani, R., Kitani, K.: MGpi: A Computational Model of Multiagent Group Perception and Interaction. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems. pp. 1196–1205 (2020)
17. Setti, F., Lanz, O., Ferrario, R., Murino, V., Cristani, M.: Multi-Scale F-Formation Discovery for Group Detection. In: 2013 IEEE International Conference on Image Processing. pp. 3547–3551. IEEE (2013)
18. Solera, F., Calderara, S., Cucchiara, R.: Socially Constrained Structural Learning for Groups Detection in Crowd. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(5), 995–1008 (2016)
19. Solera, F., Calderara, S., Cucchiara, R.: Structured Learning for Detection of Social Groups in Crowd. In: 2013 10th IEEE international conference on advanced video and signal based surveillance. pp. 7–12. IEEE (2013)
20. Solera, F., Calderara, S., Ristani, E., Tomasi, C., Cucchiara, R.: Tracking Social Groups Within and Across Cameras. IEEE Transactions on Circuits and Systems for Video Technology **27**(3), 441–453 (2016)
21. Taylor, A., Chan, D.M., Riek, L.D.: Robot-Centric Perception of Human Groups. ACM Transactions on Human-Robot Interaction (THRI) **9**(3), 1–21 (2020)
22. Turner, J.C.: Towards a Cognitive Redefinition of the Social Group. Social identity and intergroup relations **1**(2), 15–40 (1982)
23. Vascon, S., Mequanint, E.Z., Cristani, M., Hung, H., Pelillo, M., Murino, V.: A Game-Theoretic Probabilistic Approach for Detecting Conversational Groups. In: Asian conference on computer vision. pp. 658–675. Springer (2014)
24. Vázquez, M., Steinfeld, A., Hudson, S.E.: 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). pp. 3010–3017. IEEE (2015)
25. Wielemaker, J., Schrijvers, T., Triska, M., Lager, T.: SWI-Prolog. Theory and Practice of Logic Programming **12**(1-2), 67–96 (2012)
26. Wright, B., Bugajska, M., Adams, W., Lawson, E., McCurry, J.M., Trafton, J.G.: Proxemic Reasoning for Group Approach. In: 12th Intl. Conf. on Social Robotics (ICSR) (2020)
27. Zhang, L., Hung, H.: Beyond F-Formations: Determining Social Involvement in Free Standing Conversing Groups from Static Images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1086–1095 (2016)