Enabling robots to adhere to social norms by detecting F-formations

Kollakidou, Avgi; Naik, Lakshadeep; Palinko, Oskar; Bodenhagen, Leon

Published in:
2021 30th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2021

DOI:
10.1109/RO-MAN50785.2021.9515484

Publication date:
2021

Document version:
Accepted manuscript

Citation for pulished version (APA):
Kollakidou, A., Naik, L., Palinko, O., & Bodenhagen, L. (2021). Enabling robots to adhere to social norms by detecting F-formations. In *2021 30th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2021* (pp. 110-116). IEEE. https://doi.org/10.1109/RO-MAN50785.2021.9515484

Enabling Robots to Adhere to Social Norms by Detecting F-Formations

Avgi Kollakidou, Lakshadeep Naik, Oskar Palinko, Leon Bodenhagen

Abstract— **Robot navigation in environments shared with humans should take into account social structures and interactions. The identification of social groups has been a challenge for robotics as it encompasses a number of disciplines. We propose a hierarchical clustering method for grouping individuals into free standing conversational groups (FSCS), utilising their position and orientation. The proposed method is evaluated on the SALSA dataset with achieved F1 score of 0.94. The algorithm is also evaluated for scalability and implemented on a mobile robot attempting to detect social groups and engage in interaction.**

I. INTRODUCTION

As robots increasingly migrate to areas occupied by humans the need for the perception of the human social structure is growing. The use of social cues such as posture, direction of focus, and facial expressions, can aid an individual to intuitively assess the situation and act accordingly, but this remains a challenge for robots.

Humans, when gathered, tend to form groups with reserved spaces for interaction. These groups mostly maintain structures known as F-Formations [1] (see section III). The identification of such social groups and the reserved spaces can facilitate the understanding of the social context for the robot. It can be useful in cases where the robot has to avoid interrupting an ongoing human interaction by, for example, refraining from crossing the space reserved for it. It can also be useful when the robot is needed to interact in a socially acceptable way with a group as a whole while respecting its structure.

Hierarchical clustering is proposed as a basis for group identification, which is a well-established method used in statistics as well as machine learning, and implementations of it are widely available in many forms. A custom distance function that takes into account the social cues [2] available is suggested. Social signal processing classifies the behavioural cues into five categories [3]: physical appearance, gesture and posture, face and eyes behaviour, vocal behaviour and space and environment. The distance function incorporates the body orientation, location and focus points of group members, and can be easily expanded to incorporate additional features in the future. The proposed method is evaluated on a published dataset [4] as well as on a mobile robot platform.

The previous attempts for solving the problem will be considered in section II. In section III the problem formulation and the suggested clustering method is introduced and explained. Section IV introduces the evaluation metrics, presents the achieved results, the scaling capabilities of the algorithm for an increasing amount of people and finally shows the implementation of the algorithm in a real-life robotic use case. Conclusions and future work are discussed in section V.

SDU Robotics, University of Southern Denmark, Odense, Denmark
{avko,lana,ospa.lebo}@mmmi.sdu.dk

II. RELATED WORK

The interest in analysis of social interactions has seen a rise as robots move into more social environments and with it the need for automated detection of groups. Some works attempt a solution of the whole problem [5, 6], where the first attempts to capture social signals as well as perform motion reconstruction to deduce social groups and interaction using an elaborate sensor setup, with more than 500 sensors, such as VGA and RGB-D cameras providing images and point clouds of the scene, while the second proposes a joint learning framework for identification of both individual's poses and groups.

The detection and tracking of crowds is also addressed, [7, 8], where in others the problem is viewed in a more social setting to identify ways people occupy spaces or arrange themselves when interacting with each other or with technology, e.g game console or computer [9, 10]. A considerable amount of work has been done in identifying groups using videos or images for surveillance and suspect identification purposes [11, 6, 12]. The exploitation of the individual's direction, velocity and way of motion is used to identify groups that tend to move in a similar pattern [13]. The utilisation solely of images for the determination of groups using several techniques is also addressed [14, 11, 12]. In [14] head and body orientation as well as gaze direction, when possible, are taken into account to determine the Visual Focus of Attention which describes the direction of an individual's focus and utilising that to infer with who the person is possible to be interacting. The work from [11] utilises the idea of a transactional segment introduced in [15], specified as the area in front of an individual where reachability is high and effective hearing and sight is possible. The modeling of the transactional segment with a Gaussian distribution is then used to provide a graph describing the situation and the probability of any individual sharing the same o-space with another and using graph-cut algorithms for clustering the graphs. Cristiani et al. [11] propose a Hough voting strategy for finding common possible o-space centres based on the same transactional segment idea. Research in identifying groups using an array of sensors such as microphones, accelerometers, infrared signals and bluetooth is performed as well [4, 3].

Some research directly aimed at F-formation detection for mobile robotics has also been performed. A pairwise classification of individuals into F-formations and a subsequent voting for generation of larger groups is proposed in [16]. [17] uses the detection method mentioned previously in [11] to calculate approach poses for the detected groups. Finally, the work done by [18] investigates f-formation detection for automatic approach of groups by teleoperating devices.

The proposed method develops on the idea of previous works of detecting F-Formations using individuals' detected positions and orientations, proposing a new agglomerative hierarchical clustering enabling efficient computation and the transfer of the solution to a real robot platform. Social groups are detected and way-points are produced for approaching said groups in a socially acceptable way.

III. METHODS

In this section, the general definitions and terms used in the proposed solution will be introduced. In III-A, the applied clustering method will be defined, while subsection III-B defines the custom distance function utilised in the clustering. Finally, the process of detection of possible landmarks will be presented in III-C.

Free standing conversational groups (FSCGs), in most cases, are spontaneously structured in F-Formations. F-Formations are social groups that hold and maintain certain social spaces reserved only for their usage [1]. These are the o-space, defined as the common space between all members and reserved for interaction, the p-space which encompasses the o-space and includes the area occupied by the members themselves and finally the r-space which includes the immediate area surrounding the group and is reserved for individuals joining or leaving the group [19]. All spaces usually take the shape of a circle or an ellipse (fig. 1).

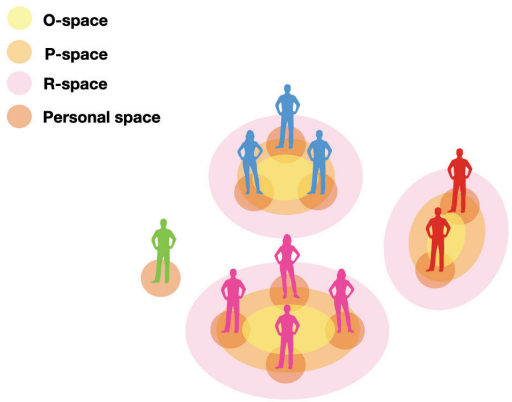


Fig. 1: F-Formations and occupied spaces in FSCGs

It is assumed that for each time frame t, the position and orientation of each individual in the room is known and described by $I_k = \{x_k, y_k, \theta_k\}$. The orientation θ, is derived by the body and head angles. The individuals are then to be divided into free-standing conversational groups. Taking the rules of proxemics into account [20], where the social space of an individual lies between 1.2 m and 3.6 m from its person, an individual is potentially interacting with any person that occupies that same space. Simplifying the premise, we assume that two individuals can be interacting if they have a distance of less than 3.6 m between them. FSCGs, on which this method is focused, usually have 2-7 members [1] and that can therefore be set as a limitation for the resulting groups. Finally, FSCGs are classified according to the focus of the group's individuals as: common-focused, when the focus lies within the group itself; jointly-focused, when all members are focused on a common point; and unfocused groups, e.g. waiting in line at a cafeteria. The method focuses on the first two, since the individuals in the third class don't participate in active interaction and are at the moment not socially relevant.

A. Clustering

A hierarchical clustering with a custom distance function is used to identify gatherings and group individuals. Specifically, the agglomerative type of the hierarchical clustering is applied, where, initially, each individual is considered as a separate cluster. Then, with every iteration the distance between each pair of clusters is calculated and the pair with minimal distance is fused into a new cluster. The method used in this case is single linkage, where we find the minimum distance between all individuals i in cluster U and individuals j in cluster V:

$$D(U, V) = \min_{i \in U, j \in V} dist(U[i], V[j]) \tag{1}$$

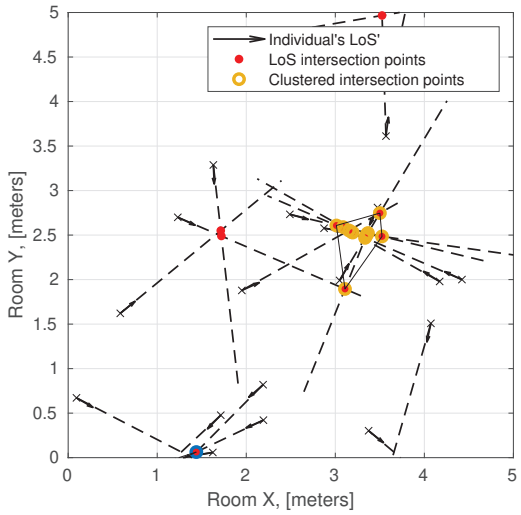
The process is then repeated until a single cluster remains or a predefined criterion is fulfilled. In this case, the criterion is set as the maximum distance allowed between cluster members. Thus, the iterations come to an end when no other clusters can be merged without violating the maximum distance criterion of $d = 3m$, which was chosen according to the proxemics distances [20]. The clustering result can be seen in fig. 2b for a simulated, randomised setting. All clustered groups are bound with a convex hull and share the same colour for visualisation.

B. Custom Distance Function

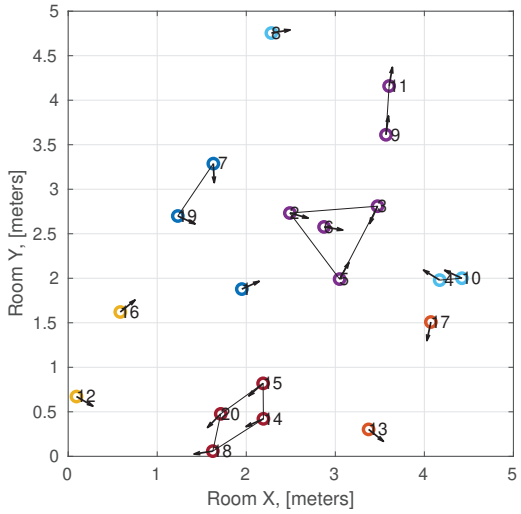
In most cases the Euclidean or Manhattan distance function is used, but as the body orientation as well as the position of individuals is to be taken into account, a custom distance function is needed. The position of each individual is described by P_k, and we define the directional unit vector of the individual's orientation as o_k. The suggested definition of the distance $d_{i,j}$ between persons i and j is as follows:

$$d_{i,j} = |P_i - P_j| \cdot \phi_{i,j} \cdot \lambda_{i,j} \tag{2}$$

Two variable coefficients are then added to the formula to account for social conventions, ϕ, the common field of view incentive coefficient, and λ, a shared focus point incentive.



(a) Clustering of LoS intersection points - Identification of Landmarks.



(b) Individuals divided into groups

Fig. 2: Simulation of randomly generated individuals. Positions and orientations displayed as arrows. Fig. 2a shows the Line of Sight (LoS) clustering for identification of landmarks. Fig. 2b shows the grouped individuals of the same setting.

The common field of view incentive coefficient, ϕ, rewards individuals that are facing towards each other and similarly penalises pairs that look away:

$$\phi_{i,j} = 1 - \kappa \cdot (\alpha_i + \alpha_j) \tag{3}$$

where α_x is a projection of the unit vector of the individual's orientation on the normalised vector between the two individuals: $\alpha_i = o_i \cdot \frac{P_j - P_i}{|P_j - P_i|}$ and equivalently $\alpha_j = o_j \cdot \frac{P_i - P_j}{|P_i - P_j|}$. Setting the variable κ, with the arbitrary value of $\kappa = 0.25$, in the case where the two are looking directly at each other $\alpha_i = \alpha_j = 1$ and therefore $\phi = 0.5$, shrinking the virtual distance between the two. Oppositely, when their orientations have a difference of $180°$, $\phi = 1.5$, increasing the distance and therefore lowering the chances of the individuals to be clustered together. Parameter κ is a tunable parameter with values $0 < \kappa < 0.5$ and can change according to the environment.

C. Landmark Detection

FSCGs can be either jointly focused or commonly focused, as mentioned before. Commonly focused groups are more troublesome to detect as they do not necessarily face each other, facing instead a common focus point, and the common field of view incentive has no, or little, effect on them. An attempt of identifying such focus points is then pursued. Each individual is assigned a landmark l_k according to their line of sight and coefficient λ is defined as:

$$\lambda_{i,j} = \begin{cases} \lambda_0 & \text{if } l_i = l_j \\ 1 & \text{if } l_i \neq l_j \end{cases} \tag{4}$$

The coefficient λ can again be tuned according to the scenario. If the environment is known to generate common focused groups, e.g. a cinema, then a smaller value of $0 < \lambda < 1$ can be used. An area can be declared a landmark if a specific number of people, or more, are directed at it. This is accomplished by observing the line of sight (LoS) of each individual and applying a separate clustering procedure on their pairwise intersections.

The LoS is defined as a closed line segment coincident with the individual's orientation, with the end point located at the distance of 3 m from the individual and starting point an offset of 0.3 m in the same direction (fig. 2a). Both the offset and the end of the line of sight values were chosen taking into account the proxemics values of personal space and social space, respectively [20].

A search for intersections between the line segments for each pair of individuals is then carried out. If one individual's LoS intersects with more than one, all intersection points are registered. These points are then clustered using a separate hierarchical clustering and the Euclidean distance between them. A threshold n is set for the minimum numbers of points a cluster must encompass to be declared a landmark. The value of n can depend on the dimensions of the observed environment, the number and density of the tracked individuals within the space and the different social occasions. If a landmark is identified, each individual whose LoS produced an intersection point in the selected cluster is assigned with the landmark's ID. All others are assigned a dummy landmark ID. Thus individuals that share the same landmark as a focus point, have a higher chance of getting clustered together. In fig. 2a, the process of LoS clustering can be seen. A LoS cluster has been detected and all affiliated individuals (marked with yellow) are now assigned with the landmark ID.

IV. RESULTS

In this section, the methods used for the algorithm's evaluation will be presented. Firstly, the evaluation on a publicly

available dataset and the achieved results will be shown in IV-A. The computational performance of the algorithm on an increasing number of people is evaluated in IV-B and finally, results of the algorithm's application on a mobile robot platform attempting to interact with social groups will be presented and discussed in IV-C.

A. Group Detection Accuracy

The proposed method is evaluated on the SALSA dataset [4]. The dataset offers annotated frames from four cameras situated on the four corners of a room. Two different social occasions are recorded in this way: a conference poster discussion session and a cocktail party. For both, positions and orientations of 18 participants were constantly tracked. An additional annotation provided by psychologists indicating which individuals belong to social groups at every frame, is provided with the dataset.

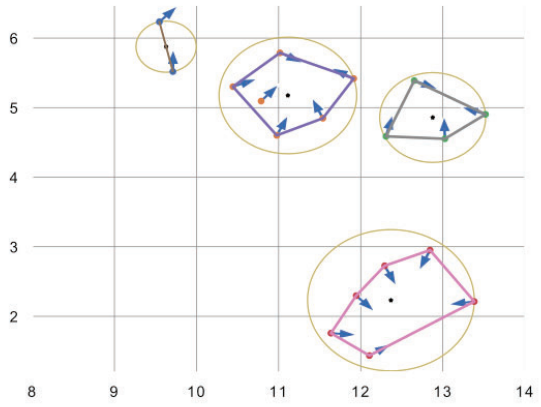
The parameters of the algorithm were set to the following values: common field of view incentive $\kappa = 0.4$, length of line of sight $LoS = 3$, threshold for landmark declaration $n = 3$ and the shared focus point incentive $\lambda_0 = 0.8$. The parameters were found to work with several group formations and distances. The parameters can be adapted to individual scenarios, for example when groups are expected to be spread, longer distance is allowed within a cluster, or in scenarios where larger amount of people are expected to be focused on a single landmark, e.g. cinema. The threshold landmark n can be configured dynamically as well, as a function of the number of people present in the scene.

For each time frame t, the 18 individuals are clustered using the proposed method. The detected groups are then compared with the provided ground truth (fig. 3). The figures show instances where all groups were detected correctly (fig. 3a) and the instance with the worst performance (fig. 3b) for the specific dataset sequence.

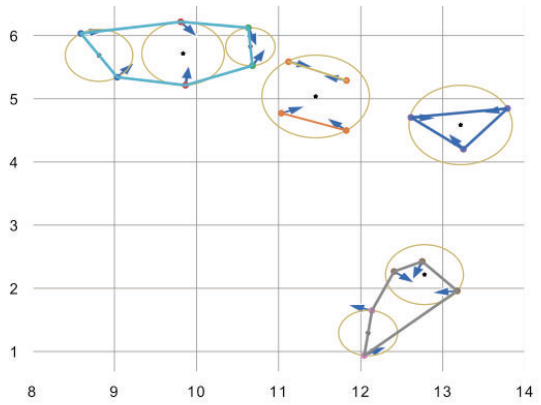
In order to quantitatively evaluate the results, two evaluation methods were used. Firstly, the known confusion matrix and classification accuracy where all groups are regarded as classes and an individual is regarded as correctly classified if they were assigned to the correct group.

Usually, a confusion matrix obtained from clustering has its columns ordered arbitrarily and the cluster-group correspondence is not established in it. In such case, to evaluate the results, columns of the confusion matrix should be reordered so that the best matching between the clusters and the ground truth is reached. Such ordering was obtained by solving a *linear assignment problem* considering the confusion matrix as its input cost map.

The results achieved for both dataset instances with the first evaluation method are promising. 92% of the individuals for the poster session and 85% for the cocktail party are assigned to the correct group. The dataset includes frames with somewhat problematic ground truth or difficult to cluster results e.g. frames of people leaving or in the process of entering a group. These were found to be the most difficult to detect and caused part of the wrong assignments.



(a) Clustering example. Best case performance.



(b) Clustering example. Worst case performance.

Fig. 3: Clustering of individuals, positions and orientations shown by the arrows. Circles: detected groups, with the equivalent o-space centres; convex hulls: ground truth annotations provided by the dataset.

Additionally, the evaluation method proposed in [12] and [21] will be used, according to which a group is considered correctly detected by the algorithm if the two following conditions are met:

$$CC \geq T \cdot G \qquad WC \leq G \cdot (1 - T) \tag{5}$$

where CC is the number of individuals that were correctly clustered into this group, WC is the number of individuals that were wrongly included in the cluster, G is the cardinality of the group in the ground truth and $0 < T \leq 1$ is the tolerance threshold by which the success of the algorithm is judged.

In this case, the values for TP, FP, and FN refer, respectively, to correctly detected groups (i.e., both conditions in (5) are met), hallucinated groups (clusters that were not assigned to any of the true groups) and missed groups, taking into account the conditions before. The precision, recall and

F1 (eqs. (6) and (7)) - the weighted average of precision and recall - values are then calculated for varying values of T.

$$precision = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN} \tag{6}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{7}$$

For the evaluation threshold the two main values of $T = 1$ and $T = 2/3$ were tested as proposed, as well as the variation of T between $0.5 \leq T \leq 1$ with 100 increments (fig. 4). Area under the F_1 curve (fig. 4), normalised to the interval length, is $\alpha = 0.84$.

The table below shows the values obtained with the two threshold values.

Dataset instance	T	precision	recall	F_1
Poster Session	2/3	0.98	0.90	0.94
	1	0.81	0.83	0.82
Cocktail Party	2/3	0.93	0.85	0.89
	1	0.66	0.68	0.67

TABLE I: Achieved evaluation values

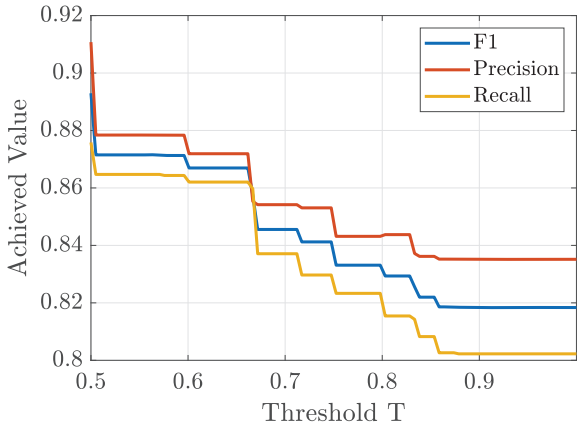


Fig. 4: Evolution of Precision, Recall and F1 value as a function of the threshold T (poster session sequence - SALSA dataset)

B. Algorithm Scalability

In order to evaluate the algorithm's computational performance in a possible higher scale, simulated randomised scenes with a varying number of individuals were clustered and the elapsed time calculated. The amount of individuals varied from 1 to 100. The evolution of required time was recorded and averaged over 100 random simulations. The accuracy of the algorithm was here of no relevance, since the purpose was the performance assessment, and was not evaluated. The calculations were carried out on a laptop equipped with an Intel® Core™ i7-8665U CPU @ 1.90GHz × 8 processor and 8GB RAM. As can be seen in fig. 5, 100 people can be assigned to groups within $0.15s$ while 10-20 people (which would be a more realistic scenario for robotic applications) could be grouped in less than $0.01s$. This shows that in both situations the grouping information can be published in a higher frequency than the one used in the mainstream navigation stack for mobile robots thus allowing real-time detection of groups.

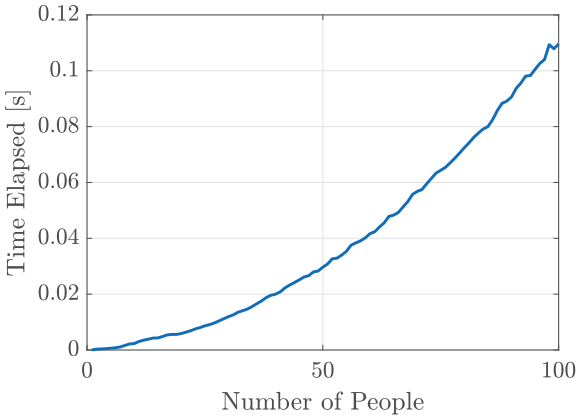


Fig. 5: Time required for clustering in respect to the amount of people present.

C. Real world experiments on robot

Aside from the experiments above, real-world experiments were also conducted to test the performance of the algorithm on noisy inputs. The experiment setup consisted of a mobile robot [22] with a 3D human pose estimation capability [23] and 9 humans standing around the robot in different F-formations in a $70m^2$ area (fig. 6). The robot's position was constant, only rotating in place for image acquisition. The experiment consisted of 10 different scenarios. In each scenario, random people were asked to discuss with each other, thus resulting in multiple conversational groups. The subjects were not introduced to the F-Formations or instructed on the way they should place themselves, rather were left free to decide that themselves when they were presented with the other members of their group, allowing for more natural formations. The structure and members of the group were changed in each scenario. A poster was placed in a part of the room and introduced to the participants, to enable the possible use of it as a focus point, which occurred in some of the scenarios. The robot's task was to identify the conversational groups among the detected people.

Fig. 6: Robot and people conversing in different F-formations during one of the real world experiment scenario

	T	precision	recall	F_1
Ground Truth	1/2	0.81	0.75	0.78
Detectable	1/2	0.84	0.94	0.88

TABLE II: Values achieved on experiments.Ground truth values are considered against the groups deduced at the time of the experiment and detectable groups take into account only the individuals present in the robot's detection

Since the robot camera has a limited field of view, it can detect only a limited number of people in a single frame. To improve the chances of detecting all the people in the environment, the robot was programmed to capture three images on its 3 sides (front, left and right) separated by 45-degree turns. In some cases, a few people were visible in multiple frames due to overlap in the robots field of view. Such detections were filtered based on proximity distance and the detections with higher detection confidence were selected as input for the grouping algorithm.

The experiment consisted of 10 different scenarios with different group members and different group sizes. Group sizes ranged from individuals to 4-person groups. In each scenario, two trials were conducted thus resulting in a total of 20 evaluations. The same evaluation strategy was used for evaluating the results as described in section IV-A. Table II presents the results of the experiment. The first row describes the performance of the algorithms w.r.t the actual ground truth, the people present in the groups. The second row describes the performance of the algorithm w.r.t the detected people, only considering for evaluation the individuals that were indeed detected by the robot for the current trial and disregarding non-detected subjects. The algorithm showed promising results in both cases, with a substantial amount of people being clustered correctly. However, it should be noted that good detections significantly improve the overall performance.

With the deduced social groups, the mobile robot is able to determine socially acceptable approach points in the case where interaction with the group members is needed. This is done by determining the p-space of said group as well as locating a position in the space where the approach would be optimal. This is inferred by employing the Voronoi lines, thus offering a point, as distant as possible from the two members closest to the robot. The point is then projected to be within the p-space, ensuring that the o-space is not invaded and no interference with the ongoing interaction occurs. The calculated approach points can be seen in fig. 7 as well as the updated costmap, to enable socially aware motion planning. For the purpose of this experiment, the robot only calculated the approach points, but did not navigate towards it.

V. Conclusion

We consider the performance of the proposed method on the SALSA dataset consistent and satisfactory. The method runs in real time on an on-board computer of a mobile robot, showing the deployability of the method in real life applications. The algorithm performs efficiently when faced with a large amount of individuals, proving the scalability of the solution.

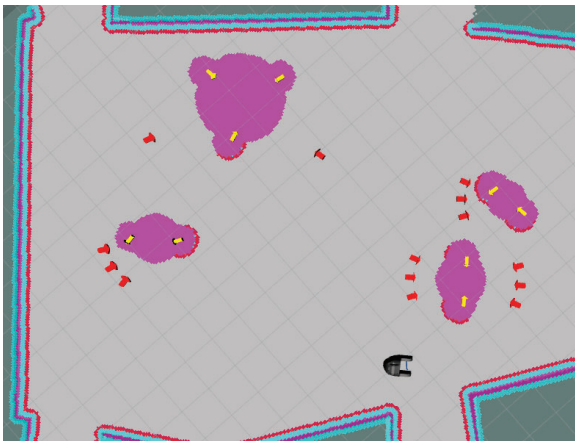


Fig. 7: Rviz visualisation of detected groups and planned approach points (yellow arrows: detected people, red arrows: planned approach points)

The method's parameters have clear roles in the way it functions and a proven positive effect on the clustering results. The method can be tuned to prioritise or ignore certain social interactions between individuals. The custom distance function can be modified in the future to include other social cues, such as gaze direction, hand gestures or environment features.

The results of the clustering are to be incorporated in the future for the implementation of socially aware navigation.

Acknowledgment

This research was supported by the project Health-CAT, funded by the European Fund for regional development.

References

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

[2] N. Ambady and R. Rosenthal. "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis." In: *Psychological bulletin* 111.2 (1992), p. 256.

[3] A. Vinciarelli, M. Pantic, and H. Bourlard. "Social signal processing: Survey of an emerging domain". In: *Image and vision computing* 27.12 (2009), pp. 1743–1759.

[4] X. Alameda-Pineda et al. "Salsa: A novel dataset for multimodal group behavior analysis". In: *IEEE transactions on pattern analysis and machine intelligence* 38.8 (2015), pp. 1707–1720.

[5] H. Joo et al. "Panoptic studio: A massively multiview system for social interaction capture". In: *IEEE transactions on pattern analysis and machine intelligence* 41.1 (2017), pp. 190–204.

[6] E. Ricci, J. Varadarajan, R. Subramanian, S. Rota Bulo, N. Ahuja, and O. Lanz. "Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 4660–4668.

[7] B. Zhou, X. Tang, and X. Wang. "Measurability of Crowd Collectiveness in Dynamic Scenes". In: ().

[8] B. Krausz and C. Bauckhage. "Loveparade 2010: Automatic video analysis of a crowd disaster". In: *Computer Vision and Image Understanding* 116.3 (2012), pp. 307–319.

[9] M. Jungmann, R. Cox, and G. Fitzpatrick. "Spatial play effects in a tangible game with an f-formation of multiple players". In: *Proceedings of the Fifteenth Australasian User Interface Conference-Volume 150*. 2014, pp. 57–66.

[10] 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.

[11] F. Setti, C. Russell, C. Bassetti, and M. Cristani. "F-formation detection: Individuating free-standing conversational groups in images". In: *PloS one* 10.5 (2015), e0123783.

[12] M. Cristani et al. "Social interaction discovery by statistical analysis of F-formations." In: *BMVC*. Vol. 2. 2011, p. 4.

[13] R. Mazzon, F. Poiesi, and A. Cavallaro. "Detection and tracking of groups in crowd". In: *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE. 2013, pp. 202–207.

[14] 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". In: *Expert Systems* 30.2 (2013), pp. 115–127.

[15] T. M. Ciolek and A. Kendon. "Environment and the spatial arrangement of conversational encounters". In: *Sociological Inquiry* 50.3-4 (1980), pp. 237–271.

[16] H. Hedayati, D. Szafir, and S. Andrist. "Recognizing f-formations in the open world". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 558–559.

[17] R. Livramento, J. Avelino, and P. Moreno. "Natural Data-driven Approaching Behaviors of Humanoid Mobile Robots for F-Formations". In: *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. IEEE. 2020, pp. 338–344.

[18] S. Krishna. "Join the Group Formations using Social Cues in Social Robots". In: *17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10-15, 2018*. Association for Computing Machinery (ACM). 2018, pp. 1766–1767.

[19] T. M. Ciolek. "The proxemics lexicon: A first approximation". In: *Journal of Nonverbal Behavior* 8.1 (1983), pp. 55–79.

[20] E. T. Hall. *The hidden dimension*. 1966.

[21] F. Setti, H. Hung, and M. Cristani. "Group detection in still images by F-formation modeling: A comparative study". In: *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE. 2013, pp. 1–4.

[22] W. K. Juel et al. "Smooth robot: Design for a novel modular welfare robot". In: *Journal of Intelligent & Robotic Systems* 98.1 (2020), pp. 19–37.

[23] W. K. Juel, F. Haarslev, N. Krüger, and L. Bodenhagen. "An Integrated Object Detection and Tracking Framework for Mobile Robots". In: *17th International Conference on Informatics in Control, Automation and Robotics (ICINCO) International Conference Informatics in Control, Automation and Robotics*. SCITEPRESS Digital Library. 2020, pp. 513–520.