Robocentric Conversational Group Discovery

Viktor Schmuck, Tingran Sheng and Oya Celiktutan

***Abstract*— Detecting people interacting and conversing with each other is essential to equipping social robots with autonomous navigation and service capabilities in crowded social scenes. In this paper, we introduced a method for unsupervised conversational group detection in images captured from a mobile robot's perspective. To this end, we collected a novel dataset called Robocentric Indoor Crowd Analysis (RICA). The RICA dataset features over 100,000 RGB, depth, and wide-angle camera images as well as LIDAR readings, recorded during a social event where the robot navigated between participants and captured interactions among groups using its on-board sensors. Using the RICA dataset, we implemented an unsupervised group detection method based on agglomerative hierarchical clustering. Our results show that incorporating the depth modality and using normalised features in the clustering algorithm improved group detection accuracy by a margin of 3% on average.**

I. Introduction

As robots are becoming progressively more widespread in our society, it is getting more important for them to take human-aware actions with full autonomy in dynamic human environments. Therefore, crowded social scene analysis, detecting people and their interactions with each other, predicting their actions and intentions plays a key role within this context. In service robotics applications, scene analysis enables robots to safely navigate in indoor spaces such as museums or airports, approach groups or individuals, and assist them in performing their tasks, or in achieving their goals through human-robot interaction. Such tasks require a mobile robot to have sufficient understanding of people and their position in different areas, as well as that of the social dynamics in the scene.

The research conducted in the past decade on crowded social scene analysis and group detection shows promising results as it utilises the concept of F-formations to determine interaction spaces [1]. Most approaches have relied on head and/or body posture detection to build models [2], based on top-down or bird-eye viewpoint images [3]–[8]. However, crowded social scene analysis from a mobile robot's perspective has not been thoroughly explored, which brings about a long list of challenges, including a narrower field of view, dynamic camera, non-ideal illumination conditions, and noise resulting from ego- or robocentric view.

This paper introduces a novel dataset called Robocentric Indoor Crowd Analysis (RICA). As shown in Fig. 1, the RICA dataset comprises of recordings from a social event where a mobile robot (i.e., Human Support Robot - HSR from Toyota Motor Europe) navigated between participants and captured the interactions among groups using its onboard RGB, depth and LIDAR sensors. The RICA dataset is annotated at both group-level (i.e., with respect to conversational groups) and individual-level (i.e., with respect to whether an individual belongs to a group). Using the RICA dataset, we focus on the problem of unsupervised group detection. Given a crowded scene image, we first extract a set of features from both RGB and depth modalities, and then use an Agglomerative Hierarchical Clustering (AHC) algorithm to identify any (unknown) number of groups. Our experimental results show that our multimodal approach improved the error rate from 1.03 to 0.98 in terms of Mean Average Error and from 1.21 to 1.15 in terms of Root Mean Square Error as compared to the state-of-the-art approach.

Fig. 1. Sample images captured by (a-b) RGB-D camera and (c-d) Wide-angle camera.

II. Related Work

This section summarises the related work from two perspectives, namely unsupervised group detection approaches based on top-view (bird-eye view) vision systems and publicly available robocentric datasets for crowd analysis.

A. Unsupervised Group Detection

Many efforts have been made on unsupervised group detection, but exclusively focusing on top-view or bird-eye view images captured via static cameras. Bazzani et al. [3] proposed DEcentralizEd Particle filtER for Joint Individual-Group Tracking (DEEPER-JIGT), where they initially identify groups based on the distribution and trajectory similarities of individuals, and then jointly track individuals and groups, recognising merging and splitting behaviour to update the groups. A line of works investigated group

*The authors are with the Centre for Robotics Research, Department of Engineering; King's College London, London, WC2R 2LS United Kingdom; {viktor.schmuck; tingran.sheng; oya.celiktutan}@kcl.ac.uk

detection in crowded videos by utilising agglomerative clustering methods [4], or social force modelling [5]. Chandran et al. [6] proposed a Non-recursive Motion Similarity Clustering algorithm that did not calculate trajectories or social forces, but motion similarities defined by distance, speed, and direction of motion. Recently, Wang et al. [9] also used akin motion similarity descriptors in conjunction with a Self-weighted Multi-view Clustering method. While these features were shown to be useful in top-view settings, they cannot be reliably extracted from the robot's viewpoint due to the height of the robot. Also, the robot's self-motion when moving and noise resulting from its sensors introduce further errors in estimating the speed of the individuals. During the analysis and navigation of crowded spaces, robots need to make fast decisions, which requires online methods that can deliver predictions from unsegmented data in a continuous stream. However, the aforementioned methods do not perform online group detection as they need sequences of images to calculate trajectories, and hence are unable to deliver the detected groups entirely on-the-fly.

To achieve an online solution, Chen et al. [8] proposed an Anchor-based Manifold Ranking method for group detection in single images. They used a small set of consecutive frames to identify individuals as anchor points within a group. Then these anchor points (centroids) were used in manifold ranking to assign the rest of the individuals to the groups in each frame. Japar et al. [10] also proposed a method based on a single image. They first detected faces with the TinyFace detector [11], then used bounding box corners and centroid coordinates as feature vectors with an array of linkage algorithms to perform Agglomerative Hierarchical Clustering [4]. Since the number of groups were within a frame was unknown, they calculated the Davies-Bouldin index [12] for each possible number of groups (K) ranging from 1 to the number of clusters detected, and then selected the K value giving the lowest score to determine the number of groups in an unsupervised manner.

B. Available Robocentric Datasets

Despite the growing need, thus far there has been only one publicly available dataset for studying social navigation and crowded social scene analysis from a robot's perspective. The Jack Rabbot Dataset and Benchmark (JRDB) [13] provides a large set of recordings containing 2D and 3D information in 360 degrees around the moving robot as well as bounding box annotations of people. However, it is not entirely applicable to robocentric indoor crowd analysis for several reasons. Most importantly, the captured environments are in general not too crowded, 1 person per 3 square meters or denser, and there is no annotation available for detecting conversational groups. Instead, this dataset comprises of recordings in both indoor and outdoor areas, usually with people queuing, walking in or out of rooms alone or in small groups.

C. Our Work

As highlighted by Taylor and Riek [14], the techniques summarised in Section II-A do not keep a robotic context in mind, as they often do not consider the unpredictability of human spaces. Moreover, they do not deal with the different types of noise introduced by the robot's sensors and movement [15], nor do they approach the problem from a robot's point-of-view. This makes the previously proposed solutions [3]–[8] less accurate when applied to an egocentric view. There have also been some efforts introducing egocentric datasets for social interaction analysis [16]–[19]. However, their egocentric data was either recorded from a distance with a static camera facing in a single direction, or the camera wearer was often already part of the conversational group [19].

Our work addresses the gaps highlighted above in several aspects. First, we collected a novel dataset called Robocentric Indoor Crowd Analysis Dataset (RICA), which features multimodal recordings of a social event attended by over 50 participants, which were captured from the viewpoint of a moving robot (i.e., Human Support Robot (HSR) from Toyota Research Europe [20]). In comparison to the JRDB dataset [13], our dataset was acquired with less high-end sensors and in indoor areas only. Nevertheless, in our dataset social scene images were denser and were annotated to enable human detection as well as group detection. Secondly, we proposed an online, unsupervised approach to group detection based on agglomerative hierarchical clustering by building upon the method proposed by Japar et al. [10]. We further improved this method by incorporating depth information as an additional modality and performing feature normalisation, as evidenced by the results from our extensive experimental evaluation.

III. Robocentric Indoor Crowd Analysis Dataset

The Robocentric Indoor Crowd Analysis (RICA) dataset[1] was recorded during a reception-style semi-public event in an indoor environment, attended by approximately 50 participants. Participants provided written informed consent, and the data collection protocol was approved by the Ethical Committee of King's College London, United Kingdom. As a robotic platform, we used the Human Support Robot (HSR) [20], a mobile support robot designed to communicate with people and hand over objects. It has 8 degrees of freedom (DoF) for manipulation, 3 DoF of the mobile base (which is equipped with IMU and laser ranger sensors), 4 DoF of the arm, and 1 DoF of the torso lift. It also has 2 DoF of its head, which has an array of both 2D and 3D cameras as well as a microphone for input sensors. The robot recorded the event with an "ASUS Xtion PRO LIVE" – RGB-D – camera, a wide-angle camera (Nippon Chemi-Con NCM13-J-02), and a "Laser measuring range sensor (UST-20LX)" – LIDAR – sensor for over one hour.

To obtain a diverse dataset, the robot was remotely driven around at different speeds, following varying paths. Using the height and head adjustment of the robot, its cameras were raised to different elevations, and its head was set to record at a variety of tilt and roll angles. Sample snapshots

[1]For further details about the dataset, visit https://sairlab.github.io/rica/.

Fig. 2. An annotated image recorded with the RGB-D camera, showing a person (ID 21 – blue bounding box on the right-hand side) not belonging to any group, and two individuals (IDs 19-20 – red bounding boxes in the middle) belonging to group ID 57 (green bounding box in the middle), where the group formation of group ID 57 is annotated as *face-to-face*.

from the dataset can be seen in Fig. 1, where the image data was captured at a resolution of 640×480. For privacy-preserving reasons, the faces of the attendees were blurred and only distance and image data was recorded (i.e., no audio data was collected). The participants were aware of the recording taking place but were avoided and not disturbed by the navigating robot for the entire duration of the event. In addition to image and LIDAR data, we recorded IMU measurements of the robot and the joint positions of its head while moving, for example, which can be used to find correspondence between image modalities and LIDAR readings (963 samples from -2.098 to 2.098 radians per sample).

We annotated the dataset using a modified and improved version of the Actanno annotation tool [21]. We performed two types of manual annotation: (1) group-level annotation – we labelled bounding boxes enclosing the groups and assigned a unique identifier (ID) to each group per frame; (2) person-level annotation – we labelled bounding boxes enclosing individuals and assigned them to the group IDs they belonged to if any. In addition, the group-level annotation involved labelling the type of group formation (F-formation) that groups of people displayed. These F-formation types included L-arrangement, face-to-face, side-by-side, semi-circular, and rectangular [22]. In summary, all RGB-D images (a total of $40,336$ frames) were annotated at the group-level with respect to bounding boxes, group IDs, and the five types of group formation. Out of this, $8,148$ frames were further annotated at the person-level. The annotations of the remaining modalities (e.g. LIDAR readings) can be automatically derived from the labelled bounding boxes based on the timestamps and the joint positions. A sample annotated image can be seen in Fig. 2.

A. Summary Statistics of the Data

The sample size of the collected data compared to JRDB can be seen in Fig. 3. While our dataset has a lower number of image samples than JRDB, it has also been annotated on the group-level, indicating F-formations of conversational groups. Regarding these annotations, the dataset comprises of 1 to 4 groups in each frame, with an average of 1.62 groups across the entire dataset. The groups are visible for

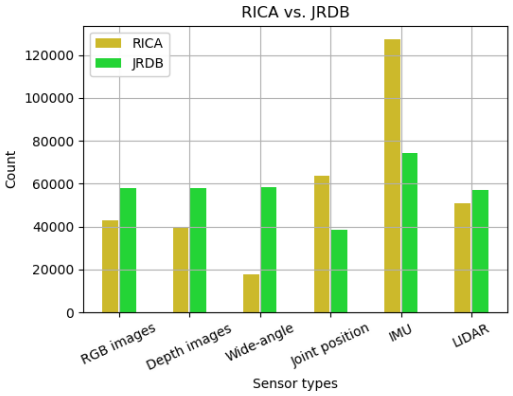

Fig. 3. Sample size comparison of the recorded modalities in RICA and JRDB.

periods ranging from a single frame to 597 frames, with an average period of 108.19 frames. In total, there are 112 distinct groups identified in the dataset. However, we did not take into account the cases where a group disappeared from the view and reappeared at a later time instant, and we considered these cases as new distinct groups.

Regarding person-level annotations, the number of people ranges from 1 to 8 with an average of 3.92 individuals per frame. Some people are not in groups, and as a result, there are 1 to 5 people with an average of 1.54 individuals in each frame who are not assigned to groups. In this paper, the group formation annotations have not been used as we focus on group detection only.

B. Benchmarking Human Detection Algorithms

We defined a series of tests to evaluate the performance of state-of-the-art human detection algorithms on our collected dataset. In particular, we test three methods on our RICA dataset, without fine-tuning: (1) Histogram of Oriented Gradients (HOG) [23] combined with non-maxima suppression (NMS); (2) MobileNet-SSD (SSD) [24] – trained on MS-COCO [25], and then fine-tuned on VOC0712 [26] – with centroid tracking, and (3) YOLO [27] – trained on MS-COCO [25]. In addition, we detected faces with the TinyFace detector (TF) [11] – trained on the WIDER-face dataset [28], which was later used as one of the input types to the unsupervised group detection method. After retrieving the bounding boxes by using all four methods (i.e., HOG, SSD, YOLO and TF), we computed their intersection over union (IOU) values against GT. However, for the TF detector, since the detected bounding boxes were much smaller than the GT bounding boxes, we computed the ratio between the area overlapping with the GT and the whole area of the detected box by TF.

The results of these comparisons are shown in Fig. 4. The best mean IOU score ($\mu = 0.64, \sigma = 0.26$) was obtained with the SSD detector, and the TF detector yielded boxes with large overlapping areas as compared to GT (area overlap $\mu = 0.88, \sigma = 0.22$). Therefore, we used the outputs of the SSD and TF detectors to implement our group detection method as described in the next section.

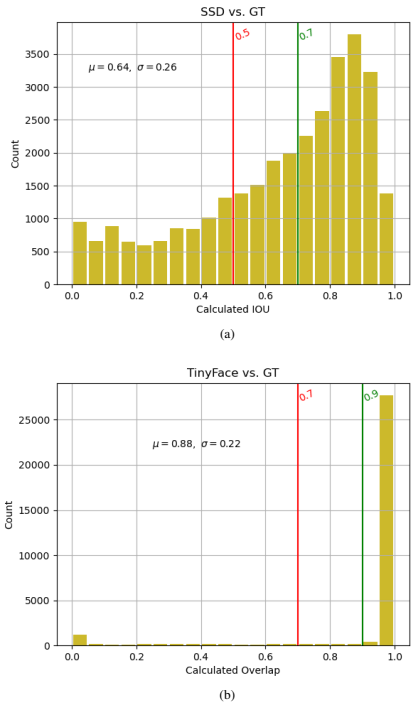

Fig. 4. Histograms of IOU values when comparing (a) GT and SSD. (d) Histogram of overlapping area values for between GT and TF. The red vertical lines show the minimum IOU and overlap scores to consider a bounding box as a True Positive detection. Green vertical lines indicate the IOU and overlap scores above which the detection is considered as successful.

IV. Unsupervised Group Detection

In this paper, we exclusively focused on the problem of unsupervised group detection and left the problem of group formation recognition as future work. In particular, we investigated the contribution of depth modality and feature normalisation by building upon the method proposed by Japar et al. [10]. The pipeline of our proposed approach was as follows. We first obtained the bounding boxes automatically and extracted a set of features describing the location and depth information. These features, or their normalised values, were then used as input to the agglomerative hierarchical clustering method to find the number of conversational groups in an image.

A. Feature Extraction

Given a single crowd image I, the problem of group detection can be defined as identifying social clusters, denoted by $c = (c_1, ..., c_k, ..., c_K)$ where $k = \{1, ..., K\}$. Differently from [10], in our case, there were individuals in the scene who did not belong to any group. K refers to the number of computed clusters as not all clusters correspond to conversational groups. Therefore, conversational groups S can be defined as $S \subseteq c$, where any $c_k \in S$ if $c_k = (c_{k1}, ..., c_{kl}, ..., c_{kL})$ and $L >= 2$, where L is the number of people forming a conversational group.

Once an individual p is detected in a crowd image I using one of the methods described in Section III-B, the individual p can be described with feature vectors A_p and B_p. Inspired by [10], we define feature vector $A_p = \{a_p^x, a_p^y, a_p^w, a_p^h\}$, where a_p^x and a_p^y are the spatial coordinates of the upper-left corner, and a_p^w and a_p^h are the width and height values of the bounding box, respectively. Similarly, $B_p = \{b_p^{cx}, b_p^{cy}\}$ is defined by calculating the centroid coordinates (b_p^{cx} and b_p^{cy}) of the bounding box. Japar et al. [10] also reported that the concatenation of the two RGB features yielded the best results, hence we define $C_p = \{a_p^x, a_p^y, a_p^w, a_p^h, b_p^{cx}, b_p^{cy}\}$, the concatenation of feature vectors A_p and B_p.

In addition to these 4-dimensional and 2-dimensional features extracted from RGB images, we designed a new feature using the depth modality. From a depth image corresponding to an RGB image, we retrieved the depth values within the bounding box for each individual. As a feature, we computed the weighted average of the depth values to take into consideration the depth sensor's noise and situations where individuals were occluded by static objects in the conversation floor, or by each other. Also, to maximise the area occupied by an individual in a box, we used 90% of the detected box areas, resulting in depth values D_{val}. The weight for each pixel D_w can be calculated by:

$$d_w^{i,j} = \sqrt{(b_p^{cx} - i)^2 + (b_p^{cy} - j)^2}, \quad \text{for all } i \text{ and } j$$
$$\text{where} \quad b_p^{cx} - \frac{a_p^w \times 0.9}{2} \leq i \leq b_p^{cx} + \frac{a_p^w \times 0.9}{2} \quad , \qquad (1)$$
$$\text{and} \quad b_p^{cy} - \frac{a_p^h \times 0.9}{2} \leq j \leq b_p^{cy} + \frac{a_p^h \times 0.9}{2} \quad .$$

Then for each individual p, the depth value d_p can be calculated by:

$$d_p(D_w, D_{val}) = \frac{\sum_{n \in j}^{N} \sum_{m \in i}^{M} D_w^{m,n} \times D_{val}^{m,n}}{M \times N} \quad , \qquad (2)$$
$$\text{where} \quad M = a_p^w \times 0.9 \text{ and } N = a_p^h \times 0.9 \quad .$$

We combined the depth features with RGB features and obtained three multimodal feature vectors, namely, $A_p^d = \{a_p^x, a_p^y, a_p^w, a_p^h, d_p\}$, $B_p^d = \{b_p^{cx}, b_p^{cy}, d_p\}$, and $C_p^d = \{a_p^x, a_p^y, a_p^w, a_p^h, b_p^{cx}, b_p^{cy}, d_p\}$. Furthermore, to investigate the effect of normalised feature inputs, we calculated the same feature vectors with their min-max normalised values, which were denoted by A'_p, B'_p and C'_p for RGB features and by A'^d_p, B'^d_p and C'^d_p for multimodal features.

B. Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering is a bottom-up approach. Given a crowd image I, each individual p represents a cluster with a single element initially. At each step of the algorithm, two clusters are merged based on similarity, until a single cluster is created. This similarity comparison is guided by different linkage approaches. In our experiments, we used the average [29] and ward [30] linkage algorithms, which were found to be the best performing methods according to the state-of-the-art results [10].

Due to the nature of unsupervised detection, the number of clusters should be detected is unknown in advance. To determine the optimal number of clusters, we used the Calinski and Harabasz Score [31] (CH score), which is a ratio calculated based on the within- and between-cluster

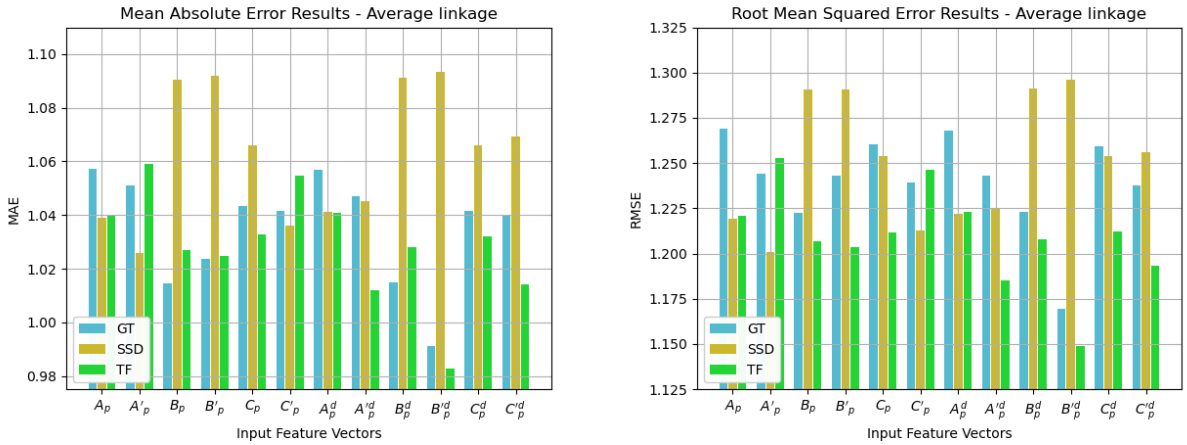

Fig. 5. Mean Average Error and Root Mean Squared Error results with Average linkage methods. $A_p = \{a_p^x, a_p^y, a_p^w, a_p^h\}$: a_p^x and a_p^y – spatial coordinates of the upper-left corner, a_p^w and a_p^h – width and height information of the box. $B_p = \{b_p^{cx}, b_p^{cy}\}$: centroid coordinates ($b_p^{cx}$ and b_p^{cy}) of the bounding box. $C_p = \{a_p^x, a_p^y, a_p^w, a_p^h, b_p^{cx}, b_p^{cy}\}$: concatenated A_p and B_p. $'$-ed terms indicate feature vectors with normalised inputs. d superscripted terms indicate added – d_p – depth feature.

dispersion. To determine the optimal number of clusters, on each iteration of the AHC method we measured the CH score. The clusters with the highest resulting score were chosen as the solution that best described the image.

V. Experimental Results

We evaluated the proposed method for unsupervised group detection on the RICA dataset. In particular, we compared the two methods for obtaining the bounding boxes (i.e., SSD and TF) with the ground-truth (GT). We systematically evaluated the three different feature types both singly and jointly, and two different linkage approaches (i.e., average linkage vs. ward linkage) for agglomerative hierarchical clustering. In Fig. 5, we presented the results in terms of Mean Average Error and Root Mean Squared Error by following [10].

The results of the comparison are given in Fig. 5. We observed that the use of average linkage and weighted linkage made no significant difference between the resulting MAE and RMSE scores, as the difference was below 1%. Therefore, we make no distinction between the two linkage types henceforth.

As for human detectors, GT and TF generated bounding boxes both yielded the smallest MAE (1.01) on average, while the generated TF bounding boxes outperformed both other methods based on the RMSE measurements (1.21). This might be due to the fact that bounding boxes are smaller, therefore their weighted average depth information is less likely to include confounding factors such as occlusions in the final feature. Moreover, it can be observed that the SSD input was outperformed by both other input types in many cases, which can be the result of inaccurate detections (see III-B), as even though it performed best out of the three tested human detectors, its mean IOU score is low.

As for the added depth modality, we present how it improved the error metrics (MAE and RMSE) for 6 feature vectors A_p, B_p, C_p, A'_p, B'_p and C'_p for GT, SSD and TF generated inputs. As illustrated by our results in Fig. 5, the added depth modality improves group detection for GT and TF generated inputs as compared to the original features (i.e., A_p, B_p, C_p) proposed by Japar et al. [10]. The best results obtained with GT bounding boxes as input are 0.99 (3% improvement) and 1.17 (4% improvement) in terms of MAE and RMSE, respectively, obtained for feature vector $B_p^{'d}$. Similarly, the best error rates for the TF generated inputs are achieved with feature vector $B_p^{'d}$, and the resulting MAE and RMSE scores are 0.98 (6% improvement) and 1.15 (5% improvement), respectively. When SSD generated bounding boxes are used, the added depth information either increases the error rates or has no effect. In other words, adding depth to the feature vectors did not improve the solution in the case of the TF generated input. As discussed earlier, this might be due to the fact that the TF generated boxes are less prone to occlusions when detected correctly, resulting in more reliable depth features.

Lastly, we observed that normalising the input features results in significant improvements when the GT and TF human detectors are used. Our results show that feature normalisation of our multimodal approach improved the error rate from 1.03 to 0.98 in terms of Mean Average Error and from 1.21 to 1.15 in terms of Root Mean Squared Error.

VI. Conclusion and Future Work

In this paper, we introduced RICA, a novel robocentric dataset for indoor crowd analysis. We used the RICA dataset to enhance and extend a state-of-the-art unsupervised group detection method by including depth information and feature normalisation. Our results showed that both techniques improved the overall accuracy of the group detection in challenging robocentric images. In addition, we compared multiple detectors to acquire human bounding boxes and showed that in most cases detecting faces only could be a better approach rather than taking into account full-body bounding boxes for group detection. Our future work will focus on developing novel methods for unsupervised online group detection and group formation recognition using the RICA dataset.

ACKNOWLEDGEMENTS

We thank Toyota Motor Research Europe for providing the Human Support Robot (HSR) as the robotic platform.

REFERENCES

[1] A. Kendon, *Conducting interaction: Patterns of behavior in focused encounters*, ser. Conducting interaction: Patterns of behavior in focused encounters. New York, NY, US: Cambridge University Press, 1990.

[2] C. Raman and H. Hung, "Towards automatic estimation of conversation floors within F-formations," *arXiv:1907.10384 [cs]*, Jul. 2019, arXiv: 1907.10384.

[3] L. Bazzani, M. Cristani, and V. Murino, "Decentralized particle filter for joint individual-group tracking," in *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1886–1893.

[4] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-Based Analysis of Small Groups in Pedestrian Crowds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 1003–1016, May 2012, ISSN: 1939-3539.

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

[6] A. Chandran, "Identifying social groups in pedestrian crowd videos," 2015.

[7] N. Elassal and J. H. Elder, "Unsupervised Crowd Counting," in *ACCV*, 2016.

[8] M. Chen, Q. Wang, and X. Li, "Anchor-based group detection in crowd scenes," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ISSN: 2379-190X, Mar. 2017, pp. 1378–1382.

[9] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting Coherent Groups in Crowd Scenes by Multiview Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 46–58, Jan. 2020, ISSN: 1939-3539.

[10] N. Japar, C. S. Chan, and V. J. Kok, "Coherent Crowd Analysis in Still Image," Sep. 2019, pp. 1–6.

[11] P. Hu and D. Ramanan, "Finding Tiny Faces," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[12] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, ISSN: 1939-3539.

[13] R. Martín-Martín, H. Rezatofighi, A. Shenoi, M. Patel, J. Gwak, N. Dass, A. Federman, P. Goebel, and S. Savarese, *JRDB: A Dataset and Benchmark for Visual Perception for Navigation in Human Environments*. 2019.

[14] A. Taylor and L. D. Riek, "Robot Perception of Human Groups in the Real World: State of the Art," en, in *2016 AAAI Fall Symposium Series*, Sep. 2016.

[15] A. Tapus, A. Bandera, R. Vazquez-Martin, and L. V. Calderita, "Perceiving the person and their interactions with the others for social robotics – A review," en, *Pattern Recognition Letters*, Cooperative and Social Robots: Understanding Human Activities and Intentions, vol. 118, pp. 3–13, Feb. 2019, ISSN: 0167-8655.

[16] A. Fathi, J. Hodgins, and J. Rehg, "Social Interactions: A First-Person Perspective," in *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1226–1233.

[17] S. Alletto, G. Serra, S. Calderara, F. Solera, and R. Cucchiara, "From Ego to Nos-Vision: Detecting Social Relationships in First-Person Views," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2014.

[18] M. Aghaei, M. Dimiccoli, and P. Radeva, *With Whom Do I Interact? Detecting Social Interactions in Egocentric Photostreams*. 2016, _eprint: 1605.04129.

[19] S. Bano, J. Zhang, and S. J. McKenna, "Finding Time Together: Detection and Classification of Focused Interaction in Egocentric Video," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, ISSN: 2473-9944, Oct. 2017, pp. 2322–2330.

[20] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase, "Development of Human Support Robot as the research platform of a domestic mobile manipulator," en, *ROBOMECH Journal*, vol. 6, no. 1, p. 4, Apr. 2019, ISSN: 2197-4225.

[21] C. Wolf, E. Lombardi, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandréa, C.-E. Bichot, C. Garcia, and B. Sankur, "Evaluation of video activity localizations integrating quality and quantity measurements," *Computer Vision and Image Understanding*, vol. 127, pp. 14–30, Oct. 2014, ISSN: 1077-3142.

[22] P. Marshall, Y. Rogers, and N. Pantidi, "Using F-formations to analyse spatial patterns of interaction in physical environments," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 2011, pp. 445–454.

[23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, ISSN: 1063-6919, vol. 1, Jun. 2005, 886–893 vol. 1.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 21–37.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," en, in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2014, pp. 740–755.

[26] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *arXiv:1506.02640 [cs]*, May 2016, arXiv: 1506.02640.

[28] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[29] R. R. Sokal, C. D. Michener, and U. of Kansas., *A statistical method for evaluating systematic relationships*, English. Lawrence, Kan.: University of Kansas, 1958.

[30] J. H. W. Jr., "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[31] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.