RICA: Robocentric Indoor Crowd Analysis Dataset

Viktor Schmuck and Oya Celiktutan
Centre for Robotics Research, Department of Engineering
King's College London, London, United Kingdom
{viktor.schmuck; oya.celiktutan}@kcl.ac.uk

https://doi.org/10.31256/Io1Sq2R

***Abstract*—In this paper, we introduce an egocentric dataset recorded from a robot's point of view (robocentric), which has been created to serve as a platform for indoor crowd analysis. The dataset features over 100,000 RGB, depth, and wide-angle camera images as well as LIDAR readings, recorded during a social gathering where the robot captured group interactions between participants using its on-board sensors. We evaluated three different human detection algorithms on our dataset to demonstrate the challenges of indoor crowd analysis from a robot's perspective.**

***Index Terms*—Indoor crowd analysis; Multisensory egocentric dataset; Group recognition**

I. Introduction

Crowd analysis can enable robots to navigate in indoor spaces, approach groups or individuals, and through human-robot interaction assist them in their tasks or in achieving their goals. The research conducted during the past decade on crowd analysis and group detection shows promising results as it utilises the concept of F-formations [1] in order to determine interaction spaces. Most approaches have relied on head and/or body posture detection to build models [2], based on top-down or a bird-eye viewpoint images.

As highlighted by Taylor and Riek [3], these techniques 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 [4], nor do they approach the problem from a robot's point-of-view, which makes them less accurate when applied to an egocentric view.

As shown in Fig. 1, to address the aforementioned gaps, we collected a novel Robocentric Indoor Crowd Analysis (RICA) Dataset [1] using Toyota's Human Support Robot (HSR) [5] as a robotic platform. In particular, we recorded a crowded, semi-public indoor event using robot's on-board cameras as well as LIDAR sensor. In comparison to the existing datasets such as the JackRabbot Dataset [6], the RICA dataset was acquired with less high-end sensors, and we annotated it to enable human detection and group recognition. In this paper, we discuss the challenges of crowd analysis from a robot's perspective and compared three benchmark human detection methods on our dataset.

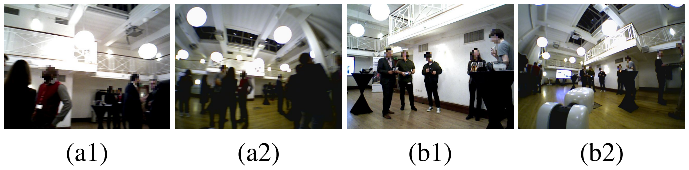
(a1) (a2) (b1) (b2)

Fig. 1. RGB-D (a1, b1) and Wide-angle camera (a2, b2) samples from two different timestamps of the RICA dataset.

TABLE I
Summary of the collected data using robot's on-board sensors compared to the relevant recordings of JRDB.

Sensor Type	Num. of Samples		Average Framerate	
	RICA	*JRDB*	*RICA*	*JRDB*
RGB camera	43,060	57,713	10.542	15.116
Depth camera	39,909	57,714	9.771	15.116
Wide-angle camera	17,877	58,313	4.377	15.273
Joint position	63,569	38,476	15.563	10.078
IMU	127,324	74,234	31.172	19.443
LIDAR	50,926	56,844	12.468	14.888

II. Robocentric Indoor Crowd Analysis Dataset

The proposed dataset was recorded during a reception-style semi-public event in an indoor environment with Toyota's Human Support Robot (HSR) [5]. 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. The dataset contains over an hour-long recording of 50 people conversing at a departmental party. Attendees were provided written informed consent, and the data collection protocol was approved by the Ethical Committee of King's College London, United Kingdom. Moreover, for privacy-preserving reasons, the face of the attendees was blurred and only distance and image data was collected.

To obtain a diverse dataset, the robot was 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. Examples of the camera's captured images can be seen in Fig. 1, where the image data was captured at a resolution of 640×480. We also recorded IMU measurements of the robot and the joint positions of its head while moving, which can be used to find correspondence between image modalities and LIDAR readings (963 samples from -2.098 to 2.098 radians per sample). The number of samples and average rate per modality are given in Table I.

We thank Toyota Motor Europe for providing the Toyota HSR robot as a development platform.

[1] The dataset will be made available at https://sairlab.github.io/rica/.

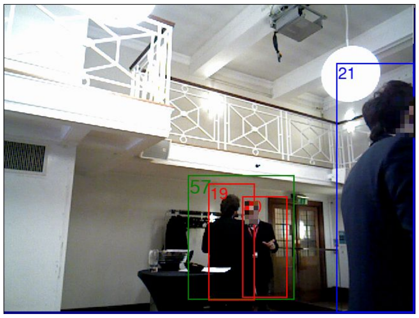

Fig. 2. An annotated image recorded with the RGB camera, showing a person (ID 21 – blue bounding box on the right hand side) not belonging to any group, and two individuals (IDs 19-20 – red bounding boxes in the middle) belonging to group ID 57 (green bounding box in the middle), where the group formation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

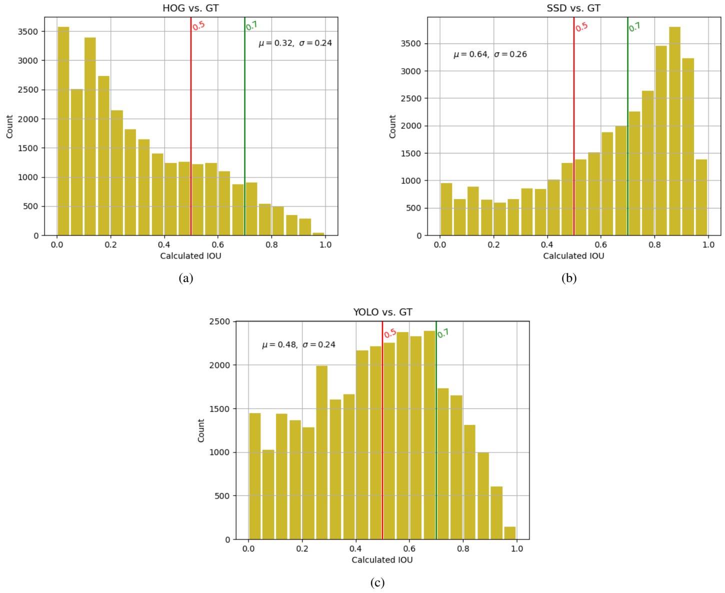

Fig. 3. Histograms of IOU values for between GT and (a) HOG; (b) SSD; and (c) YOLO. The red vertical lines show the minimum IOU and overlap scores to consider a bounding box as a True Positive detection. Green vertical lines indicate the IOU and overlap scores above which the detection is considered as successful.

We labelled the dataset by using a modified version of the Actanno annotation tool [7]. All RGB images of the dataset have been labelled at a group-level, plus identifying the group formations (i.e. L-arrangement, face-to-face, side-by-side, semi-circular, and rectangular) [8]. In addition, person-level labelling, marking people and assigning them to the identified groups, has been done for a total of $8,148$ RGB images. These person-level annotations show that in each frame there are 1 to 8 people with an average of 3.92 individuals per frame. The annotations of the remaining modalities 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. *Challenges*

The RICA dataset was collected without providing participants a script, therefore capturing the natural behaviour of attendees when a mobile support robot was navigating the event floor. The robot was driven at different speeds, on randomised paths, while its cameras were raised to different elevations and its head was held in different angles as it observed the interaction groups. Our manual inspection of the data shows that this resulted in high variation in the camera-to-subject distance (0.1-25m), and participants were often occluded by static objects or each other. It was not ensured that all participants of a single group were in the field of view of the robot and the observation length of each group was varied. The height variation introduces colour changes in the observations of the RGB-D and Wide-angle camera's images. Due to these factors, the recognition of individuals, groups, and group types from the robot's viewpoint is a challenging task, and datasets dedicated to robocentric settings are crucial to advancing the state-of-the-art.

III. Evaluation

We define a series of tests to evaluate the performance of state of the art human detection algorithms on the collected dataset. In particular, we test three methods on the RICA dataset, without fine-tuning: (1) Histogram of Oriented Gradients (HOG) [9] combined with non-maxima suppression (NMS); (2) MobileNet-SSD (SSD) [10] – trained on MS-COCO [11], and then fine-tuned on VOC0712 [12] – with centroid tracking, and (3) YOLO [13] – trained on MS-COCO [11]. After retrieving the bounding boxes with each of the human detection methods from the person-level annotated RGB images of the RICA dataset, we computed their intersection over union (IOU) values against ground truth (GT).

Even though there is minimum a single person in each frame, the HOG+NMS detector failed to detect any humans in over 11% of the images, whereas the MN-SSD and YOLO exhibited a similar, better performance – not detecting any humans in 0.7% of all frames. The results of the IOU comparisons are given in Fig. 3. The best mean IOU score ($\mu = 0.64$) was obtained with the SSD detector.

IV. Conclusion and Future Work

In this paper, we introduced a novel robocentric dataset for indoor crowd analysis, called RICA. Our preliminary analysis shows that the state-of-the-art human detectors fall short and sometimes are unable to detect any humans in the scene due to a list of challenges as summarised in Section II-A. As future work, we will investigate how we can improve human detection and tracking, e.g. by employing occlusion handling techniques in tracking [14]. Moreover, we aim to design an unsupervised approach to group detection in indoor crowded scenes – by adding modalities other than RGB image inputs, and utilising F-formations – based on the RICA dataset.

References

[1] A. Kendon, *Conducting interaction: Patterns of behavior in focused encounters*. 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.

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

[4] 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," *Pattern Recognition Letters*, Cooperative and Social Robots: Understanding Human Activities and Intentions, vol. 118, pp. 3–13, Feb. 2019.

[5] 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," *ROBOMECH Journal*, vol. 6, p. 4, Apr. 2019.

[6] 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.

[7] 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, 2014.

[8] 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.

[9] 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)*, Jun. 2005, 886–893 vol. 1.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] F. Camara, N. Bellotto, S. Cosar, D. Nathanael, M. Althoff, J. Wu, J. Ruenz, A. Dietrich, and C. W. Fox, "Pedestrian Models for Autonomous Driving Part I: Low level models, from sensing to tracking," *arXiv:2002.11669 [cs]*, Feb. 2020, arXiv: 2002.11669. [Online]. Available: http://arxiv.org/abs/2002.11669 (visited on 04/08/2020).