Workshops at 18th International Conference on Intelligent Environments (IE2022)
H.H. Alvarez Valera and M. Luštrek (Eds.)

doi:10.3233/AISE220019

Dataset to Detect Emotions from a Robot-Centric Perspective

Marco QUIROZ [a,1], Raquel PATIÑO [a], Jose DIAZ-AMADO [a,b], Yudith CARDINALE [a,c]

[a] *Electrical and Electronics Engineering Department, Universidad Católica San Pablo, Arequipa, Peru*
[b] *Electrical Engineering, Instituto Federal da Bahia, Vitoria da Conquista, Brazil*
[c] *Universidad Internacional de Valencia, Spain*

Abstract.
Social robotics is an emerging area that foster the integration of robots and humans in the same environment. With this objective, robots include capacities such as the detection of emotions in people to be able to plan their trajectory, modify their behavior, and generate a positive interaction with people based on the information analyzed. Several algorithms developed for robots to accomplish different tasks, such as people recognition, tracking, emotion detection, demonstrate empathy, need large and reliable datasets to evaluate their effectiveness and efficiency. Most existing datasets do not consider the first-person perspective from the sensory capacity of robots, but third-person perspective from out of the robot cameras. In this context, we propose an approach to create datasets with a robot-centric perspective. Based on the proposed approach, we made up a dataset with 23,222 images and 24 videos, recorded from the sensory capacity of a Pepper robot in simulated environments. This dataset is used to recognize individual and group emotions. We develop two virtual environments (a cafeteria and a museum), where there are people alone and in groups, expressing different emotions, who are then captured from the point of view of the Pepper robot. We labeled the database using the Viola-Jones algorithm for face detection, classifying individual emotions into six types: happy, neutral, sad, disgust, fear, and anger. Based on the group emotions observed by the robot, the videos were classified into three emotions: positive, negative, and neutral. To show the suitability and utility of the dataset, we train and evaluate the VGG-Face network. The efficiency achieved by this algorithm was 99% in the recognition of individual emotions and the detection of group emotions is 90.84% and 89.78% in the cafeteria and museum scenarios, respectively.

Keywords. dataset, robot-centric perspective, individual emotions, group emotions

1. Introduction

Social robots are increasingly being incorporated into crowded human spaces, such as museums, hospitals, restaurants, to offer services, perform tasks, and interact with people. Social robots should mimic the sociocognitive abilities of humans and explore behaviors

[1] Corresponding Author; E-mail:marco.quiroz@ucsp.edu.pe

to be empathetic and help with Human-Robot Interactions (HRI) [2,4]. In this sense, visual perception could provide information to understand and recognize emotions, for example through the facial expression [10,11]. According to the detected emotion and the specific situation, robots adapt their actions to display appropriate behaviors. In the recognition of individual and group emotions, it is common to use facial expressions and classify them according to the six basic human emotions: happy, neutral, sad, disgust, fear, and anger. To train the Machine Learning models, datasets generated in controlled environments with good lighting and without occlusions are used, where it is common to obtain an accuracy greater than 90%. But, the challenge is to obtain a similar result in datasets generated in uncontrolled environments where there is a greater variability of light, greater occlusion, and therefore greater challenge [6,9,13,16].

In the context of social robotics, it is also necessary to consider the robots' first-person perspective of the world. Cameras mounted on the robots' head or chassis have allowed studying the scenes from a point of view that provides robots such a first-person perspective of the world. This field of research in computer vision is known as egocentric or first-person vision [1]. The third-person camera is a device outside of the robot. The egocentric vision present advantages in comparison with the third-person camera as the robot is recording exactly what it looks in front of it, the camera movement is driven by the robot's body, and the stabilization of the image is controlled by the robot itself. Robots can use an egocentric vision to recognize emotions, navigate, or detect different objects. Developing models with this perspective makes the robot able to adapt to social groups of humans [18]. Nevertheless, most existing studies related to detect individual and group emotions are based on third-person cameras [7,8,20], but their complexity makes them not suitable for social robots with egocentric vision from their sensory capacity.

In this context, we propose an approach to create datasets with a robot-centric perspective to support the development of algorithms for emotion detection in virtual environments. Therefore, we compiled a data set with 23,222 images and 24 videos, with images classified according to the six basic emotions and videos classified as positive, negative and neutral. In order to show the suitability and usefulness of the datasets created, we trained and tested the VGG-Face network to recognize individual and group emotions.

This document is structured as follows. In Section 2, we describe studies related to the importance of datasets from an egocentric perspective. Section 3 explains the methodology for obtaining a dataset of images labeled with their respective emotion and a dataset of videos labeled with their respective emotion per frame. The experiments section explains the use of the proposed database for an algorithm for emotion detection by a Pepper robot. Section 4 shows the results obtained in the detection of individual and group emotions. Discussions about the findings are presented in Section 5. Finally, Section 6 presents the final conclusions and future research.

2. Related Work

In the context of HRI, emotion recognition has become an essential strategy to adapt the behavior of social and service robots that share spaces with humans. Depending on the emotion detected, the robot can modify its behavior or its navigation, showing a socially accepted attitude. In [14], authors mention the importance of emotion recognition for HRI.

Recent research works highlight the importance of generating datasets from an egocentric perspective and in uncontrolled environments. In [3], authors use a Bayesian network and recognition of individual facial expressions in various environments conditioned by light and temperature. This work presents an approach to estimate group emotion that leads to the calculation of the emotion of the target group. There is no dataset per se, because the method used to recognize group emotions updates its values (domain knowledge) over time to improve its performance. Also based on individual facial expressions, in [5], it is described a group emotion recognition model for an entertainment robot. The evaluation dataset is not shown, but it does conclude that the inclination of the faces affects the recognition of emotions and recommends the use of more cameras to improve performance. Both studies use their own datasets, so the comparison of results among the existing proposals is difficult, since each dataset has its own qualities. In our proposal, different datasets can be generated, according to the capabilities of the robot.

Some other studies focus on the detection of people and groups from an egocentric perspective, where the robot moves through the social environment and captures videos and images that are later manually tagged, and to track people in 3D, data is also captured from a LIDAR sensor [12,15,17,19]. These datasets were generated in uncontrolled environments, where the data include variable lighting, occlusion, and natural human behavior. These data are then manually labeled. With our approach, besides manual labeling, it is also possible to generate datasets that can be automatically labeled.

3. Datasets with Robot-centric Perspective: Our Proposal

There is an increasing research in the development of methods for the detection of individual and group emotions. Hence, competitions such as EmotiW, help in the development of this research. In spite of these advances, the work related to the detection of emotions from a robotic perspective is not very common. Consequently, as far as we know, there are few datasets containing images or videos taken by robots (i.e., from an egocentric view). To do so, we propose an approach, shown in Figure 1, to generate robot-centric datasets, that can be used in the training and validation phases of emotion recognition models, both considering individuals or group of people. This dataset creation method is supported on Robot Operating System (ROS) and Gazebo, in which real robots can be represented acting in simulated scenarios populated by people expressing different emotions. The idea is to generate a dataset composed by RGB images and another dataset conformed by videos, taken from the robot sensory capacity.

In the first stage of the methodology, the virtual environment is generated (e.g., museum, office), where individuals and groups of people are represented. For the images dataset, all the virtual characters (i.e., person representations in the environment) have the same emotion, since the idea is to have different faces but with the same emotional expression. The facial expressions of each model were generated in MakeHuman. This is followed by a recording of a video sequence with the robot's front camera. Face detection is performed on each video. The labeling of the images dataset is automatic, because all faces detected with the Viola-Jones algorithm, have the same emotion. For the videos dataset, the groups are conformed by persons with different emotional expressions, as

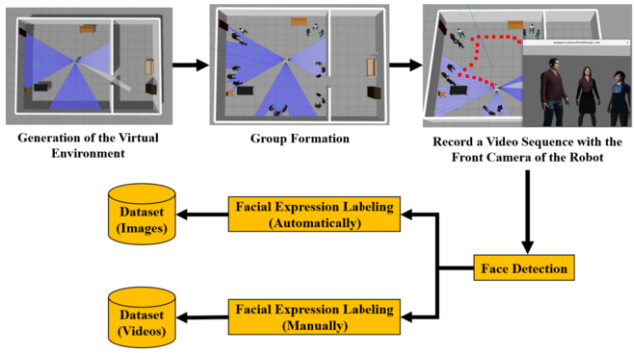

Figure 1. Applied method to create the datasets.

in a scene people can express different emotions. In this case, the labeling is performed manually. Virtual environments, models, datasets, and other ROS files are available[2].

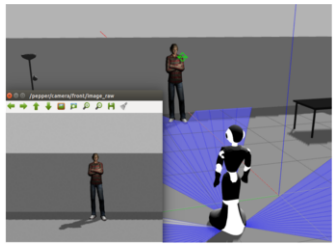

Figure 2. Image obtained by the front camera of the Pepper robot.

Following the proposed method, we generated two datasets, with the perspective of the social robot Pepper simulated in ROS/Gazebo. Pepper robot has various sensors to know its environment; in this case, only the front camera located in the upper part of the robot's head is used. This camera has a resolution of 640x480 pixels at a speed of 1-30 fps. Figure 2 shows the robot Pepper in a virtual office and the image obtained by the front camera in the robot (lower left part in Figure 2). We use ROS Kinetic and Gazebo 7 to simulate the behaviour of robots in indoor virtual environments; this version of ROS works with Ubuntu 16.04 and Python 2.7. Furthermore, to implement the proposal of this work in ROS. All simulations are performed in a desktop PC, with 16 GB RAM memory, an AMD A10 7860k 4-core CPU, and an AMD RX-570 - 4GB graphic card.

3.1. Images dataset

To generate the images dataset, a virtual office environment is used. For each emotion, we make six groups of three members, we record a video sequence with the front camera of the robot where faces are detected at different angles and directions. Then, the detected faces are automatically stored and tagged for each emotion. In total, this dataset contains 23,222 images of faces that are classified according to six emotions: happiness (4121

[2] https://github.com/marco-quiroz/Dataset-in-ROS.git

images), sadness (3895 images), anger (4113 images), surprise (3929 images), disgust (3438 images), and fear (3726 images).

3.2. Videos dataset

To generate the videos dataset, we use two virtual environments: a museum and a cafeteria (Figure 3 and Figure 4, respectively). In this case, the important issue is to form groups with people who have different emotions. In each virtual environment, we form 12 groups and record how the robot moves forward and sweeps to capture all the faces in the group. Each video consists of approximately 60 frames and is approximately 4 seconds long. Then there is manual tagging of the emotion for each frame.

Figure 3. Virtual museum used to generate the 12 videos conforming the videos dataset.

For the formation of the groups in the videos, five people were used for each emotion; that is representation of 30 people in total. These representations of people are different from those used in the creation of the images dataset. The formation of groups is carried out considering the circular formation (groups of three people or more) and the side-by-side formation (groups formed by two people). In total there are 24 videos that are used as test data to validate the emotion of each frame.

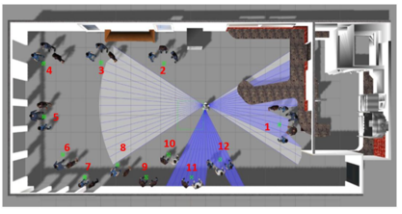

Figure 4. Virtual cafeteria used to generate the 12 videos conforming the videos dataset.

4. Results

To demonstrate the suitability and usefulness of the datasets, we train and validate a VGG-Face network to recognize emotions. Thus, we evaluate the efficiency of this trained VGG-Face to recognize individual emotions. For group emotions recognition, the individual emotions of the faces in each frame are recognized and the emotion of the frame is classified with the predominant emotion.

4.1. Individual Emotion

To validate the results of the detection of individual emotions, we have a dataset of 23,222 images, from which 82% are used as training samples and 18% as evaluation samples. Table 1 shows the number of images for each emotion and the distribution for training and test.

Table 1. Images Datset.

Emotions	Happy	Sad	Angry	Disgust	Surprise	Fear
Trainig Data	3371	3145	3363	2872	3179	2976
Test Data	750	750	750	566	750	750
Total	4121	3895	4113	3438	3929	3726

Figure 5 shows the confusion matrix of test data. The predicted labels by the model are represented on the x axis and the true labels are on the y axis. The confusion matrix shown in Figure 5 was designed with 750 validation images for each label except the disgust label with having 566 images. Finally, the validation process has only 9 incorrect images. Additionally, the maximum accuracy value during training is 0.9995 (99.95%) and the maximum accuracy value in the validation data is 0.9979 (99.79%).

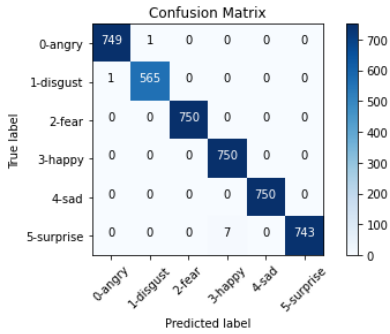

Figure 5. Confusion matrix of VGG-Face neural network.

4.2. Emotion of Videos

To validate the results of emotion detection in videos, we have a dataset of 24 videos (12 videos recorded in the virtual museum and 12 videos recorded in the virtual cafeteria). Figure 6 and Figure 7 show the emotions in each video. At the left end are the videos (Video 1, Video 2, ..., Video 12) and at the right end the average accuracy obtained is shown on each figure. The color of the points determines the emotion of the frame.

Figure 6 shows more details of the results obtained with the videos recorded in the museum. The less precise results, for example in Video 3, are due to the fact that the lighting of the virtual museum was not very good and in the case of Video 4, the detected faces did not look directly at the robot's camera. The lowest accuracy is found in Video 3, this is because the face detector (i.e., Viola-Jones model) considered other regions as faces.

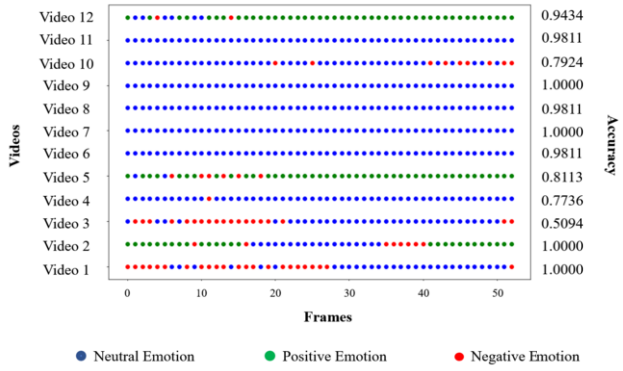

Figure 6. Results obtained for each video recorded in the Museum.

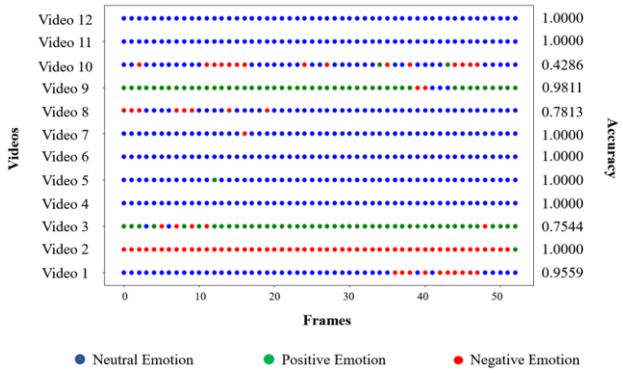

Figure 7. Results obtained for each video recorded in the Cafeteria.

Figure 7 shows the results obtained with the videos recorded in the cafeteria. The lowest precision is found in Video 10, this is because not all faces were detected in the video. In this case, the group had a neutral emotion and the undetected faces generated a wrong emotional recognition. The same happens in Video 6, but the undetected faces did not affect the result because the group had a negative emotion. The same happens in Video 6, but the undetected faces did not affect the result because the group had a negative emotion. These results demonstrate that the conditions of the environment (e.g., lighting) and the perspective of the robot may affect the final classification.

5. Discussion

Most studies developed in this context base their proposed approaches in a third-person perspective – i.e., emotions are detected through human vision or by fixed cameras in a determined place, using this information as the perspective of a robot. Results obtained in this work demonstrate that the perspective of the robot can change depending on several aspects, such as displacement, vibration, external agents, circular and angular movements, which are part of the natural process of the robot when it is moving around

the environment. In consequence, it is necessary to count with datasets that consider the sensors available in the robot (camera, Lidar, IMU, Encoder, etc.) to obtain RGB(D) images and videos from a first-person or robocentric perspective, with different wide-angle camera, joint position, etc. Thus, it will be better train, test, and validate emotion recognition models in social robotics scenarios. The approach proposed in this work to generate datasets from a robot-centric perspective is a step towards the improvement of this research area. Although it is initially applied in simulated scenarios, we believe that its implementation with real robots and real scenarios, will generate better results, especially to test real-time response models and characterization of other robots' aspects. It is certain that we will find new challenges to overcome, such as the influence of robot pose, motion, lighting, or human pose on the quality of the images and frames in the dataset and on the performance of the emotion recognition model. The consideration of these aspects will help to complement our proposal.

6. Conclusions

As reviewed in this article, most datasets to recognize emotions are generated in controlled environments and are based on a third-person perspective using fixed cameras. Therefore, we propose a methodology to generate image and video datasets with the robocentric perspective in simulated environments in ROS. To build these datasets we started from known information, in this case the emotion of the people was unique, so we collected the images of the detected faces, where the labeling is automatic. The results obtained show that lighting, displacement, and other external factors typical of the robocentric perspective influence the recognition of emotions. As future work we are going to use other Pepper robot sensors, such as the laser sensor in order to improve the labeling of emotions in dynamic environments. In addition, based on this proposed methodology, we will generate datasets using real robots, thus improving interaction between a robot and other people.

Acknowledgement

This research was funded by FONDO NACIONAL DE DESARROLLO CIENTÍFICO, TECNOLÓGICO Y DE INNOVACIÓN TECNOLÓGICA - FONDECYT as executing entity of CONCYTEC under grant agreement no. 01-2019-FONDECYT-BM-INC.INV in the project RUTAS: Robots for Urban Tourism Centers, Autonomous and Semantic-based.

References

[1] Bandini, A., and Zariffa, J. Analysis of the hands in egocentric vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* (2020).

[2] Casas, J., Gomez, N. C., Senft, E., Irfan, B., Gutiérrez, L. F., Rincón, M., Múnera, M., Belpaeme, T., and Cifuentes, C. A. Architecture for a social assistive robot in cardiac rehabilitation. In *2018 IEEE 2nd Colombian Conference on Robotics and Automation (CCRA)* (2018), Ieee, pp. 1–6.

[3] Choi, S.-G., and Cho, S.-B. Bayesian networks+ reinforcement learning: Controlling group emotion from sensory stimuli. *Neurocomputing 391* (2020), 355–364.

[4] Cooper, S., Di Fava, A., Vivas, C., Marchionni, L., and Ferro, F. Ari: The social assistive robot and companion. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (2020), IEEE, pp. 745–751.

[5] Cosentino, S., Randria, E. I., Lin, J.-Y., Pellegrini, T., Sessa, S., and Takanishi, A. Group emotion recognition strategies for entertainment robots. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2018), IEEE, pp. 813–818.

[6] Guo, X., Polania, L., Zhu, B., Boncelet, C., and Barner, K. Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2020), pp. 2921–2930.

[7] Guo, X., Polanía, L. F., and Barner, K. E. Group-level emotion recognition using deep models on image scene, faces, and skeletons. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (2017), pp. 603–608.

[8] Guo, X., Zhu, B., Polanía, L. F., Boncelet, C., and Barner, K. E. Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (2018), pp. 635–639.

[9] Huang, Y., Chen, F., Lv, S., and Wang, X. Facial expression recognition: A survey. *Symmetry 11*, 10 (2019), 1189.

[10] Liu, Z., Wu, M., Cao, W., Chen, L., Xu, J., Zhang, R., Zhou, M., and Mao, J. A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA Journal of Automatica Sinica 4*, 4 (2017), 668–676.

[11] Lopez-Rincon, A. Emotion recognition using facial expressions in children using the nao robot. In *2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)* (2019), IEEE, pp. 146–153.

[12] Martín-Martín, R., Rezatofighi, H., Shenoi, A., Patel, M., Gwak, J., Dass, N., Federman, A., Goebel, P., and Savarese, S. Jrdb: A dataset and benchmark for visual perception for navigation in human environments. *arXiv preprint arXiv:1910.11792* (2019).

[13] Mellouk, W., and Handouzi, W. Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science 175* (2020), 689–694.

[14] Mohammed, S. N., and Hassan, A. K. A. A survey on emotion recognition for human robot interaction. *Journal of computing and information technology 28*, 2 (2020), 125–146.

[15] Schmuck, V., and Celiktutan, O. Rica: Robocentric indoor crowd analysis dataset. *IMU 127*, 74,234, 31–172.

[16] Sun, M., Li, J., Feng, H., Gou, W., Shen, H., Tang, J., Yang, Y., and Ye, J. Multi-modal fusion using spatio-temporal and static features for group emotion recognition. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (2020), pp. 835–840.

[17] Taylor, A., Chan, D. M., and Riek, L. D. Robot-centric perception of human groups. *ACM Transactions on Human-Robot Interaction (THRI) 9*, 3 (2020), 1–21.

[18] Taylor, A., and Riek, L. D. Robot perception of human groups in the real world: State of the art. In *2016 AAAI Fall Symposium Series* (2016).

[19] Taylor, A., and Riek, L. D. Regroup: A robot-centric group detection and tracking system. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (2022), pp. 412–421.

[20] Xuan Dang, T., Kim, S.-H., Yang, H.-J., Lee, G.-S., and Vo, T.-H. Group-level cohesion prediction using deep learning models with a multi-stream hybrid network. In *2019 International Conference on Multimodal Interaction* (2019), pp. 572–576.