Towards Context Aware Emotion Recognition in HRI for Social Robots

Katie Powell, Sarath Kodagoda, and Teresa Vidal-Calleja
Robotics Institute, University of Technology Sydney, NSW, Australia
katie.a.powell@student.uts.edu.au

Abstract

Social robots are becoming more prevalent in our daily environments but continue to struggle communicating in human-robot interactions, often misunderstanding people and thus making the interaction uncomfortable. Many attempts have been made to improve their understanding of people and their emotions but they still lack the socio-emotional intelligence humans often use in human-human interactions. A new approach previously explored in computer science is using context emotion recognition to interpret a scene for clues to a person's emotional state. In this paper, we state that context emotion recognition will benefit the fields of human-robot interaction and social robotics. Further, we extend upon the EMOTIC model successfully adding a graphical representation of the emotion probabilities over time to the model output and with the addition of a facial feature extractor module that obtains an encouraging improvement over the original model. The algorithm was validated through data coming from two robotic platforms, namely PR2 and Pepper. The results show a promising way towards context aware emotion recognition in human-robot interactions with social robots, with 88% accuracy when comparing with 66% accuracy of the base model.

1 Introduction

Social robots are developed for the purpose of socially interacting with people in Human-Robot-Interaction (HRI) applications; from being a companion in a person's home to working at a reception desk in a hotel. Therefore, as robots become more prevalent in daily environments, it is expected that the robot's knowledge about a person's emotions can add value to the interactions. However, a robot's inability to accurately interpret and express emotions can have negative consequences for the interaction [Bartneck *et al.*, 2020].

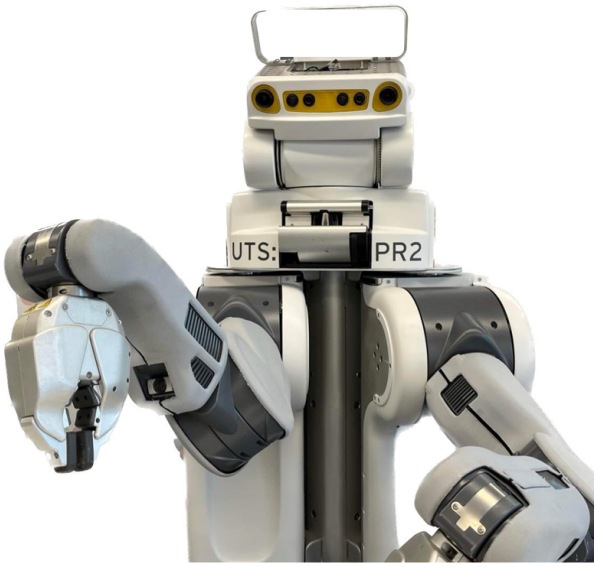

Figure 1: The PR2 robot from Clearpath Robotics.

Context emotion recognition is a field of emotion recognition recently gaining traction within psychology [Barrett *et al.*, 2011] and more recently computer science fields. It refers to the surroundings of a person including information about the environment, the activities being performed and the relationships to any people in the scene, which help determine the emotions of that person [Kosti *et al.*, 2017] [Kosti *et al.*, 2019].

HRI and social robots are lacking in effective emotion recognition technology which is further hindered when the social robot is put into real situations [Bartneck *et al.*, 2020]. As context emotion recognition has yet to be applied in HRI, there is an opportunity for new research to fill this gap. Therefore, we propose the suitability of context emotion recognition in the field of HRI for use with social robots. In this paper we will discuss and present our contributions to the EMOTIC model and its

application on a social robot in HRI.

2 Related Works

2.1 EMOTIC and Context Emotion Recognition

The EMOTIC model and dataset developed by Kosti et al. was the first context emotion recognition work in the computer science field [Kosti *et al.*, 2017]. They argue that the usually discarded context of a scene actually holds important clues to the emotional state of a person that common emotion recognition methods including facial are not able to interpret. The dataset contains 23,571 images, including 34,320 people annotated for 26 emotion categories and the discrete categories of Valence, Arousal, Dominance (VAD). Thus, it exceeds the number of emotions recognised in previous emotion recognition models, which mainly use the 6 basic emotions [Ekman, 1992]. The model is a convolutional neural network model for a multi-class multi-label problem that outputs a list of recognised emotions and a VAD score. The authors obtained an average precision score (AP) of 28.33 [Kosti *et al.*, 2017], providing a benchmark for further improvements. Therefore, the EMOTIC model and dataset have become a base for continued context emotion recognition research within the computer science field.

Since the introduction of context emotion recognition into computer science by Kosti et al., other researchers have agreed on its benefits to human-computer interaction applications and have contributed their own extensions to the idea as proof. Similar to other unimodal emotion recognition methods, context emotion recognition by itself won't provide an accurate full picture of a person's emotional state. Therefore, a multimodal approach is preferred among researchers including combinations of context and body [Kosti *et al.*, 2019], [Huang *et al.*, 2021], [Zhang *et al.*, 2019], context and face [Lee *et al.*, 2019], and context, body and face [Mittal *et al.*, 2020]. Currently, there are minimal datasets suited for context emotion recognition. However, both image [Kosti *et al.*, 2019], [Yang *et al.*, 2022] and video [Mittal *et al.*, 2020], [Huang *et al.*, 2021] datasets can be created to train these systems allowing for a wider range of applications. These datasets contain scenes of real life including public places such as bus stops, hospital entrances, and shopping centres [Mittal *et al.*, 2020], [Huang *et al.*, 2021]. This 'in the wild data' is beneficial when the intending application will also contain similar scenes, hopefully improving performance of the emotion recognition system. While EMOTIC and context emotion recognition is now deemed important in human-computer interaction applications, little is known about its effectiveness in HRI and for social robots.

2.2 Social Robots in HRI

In HRI, the robot's design is an important aspect of creating a comfortable and natural interaction between a robot and a human. The robot, in this case a social robot, should be specifically chosen to accommodate this rule. For example, research into the focal point, most often a face, determined that is where the robot's responsiveness will be expected during an interaction [Zaballa *et al.*, 2021]. Depending on the research question, robots of a humanoid form, an animal-like form, and machine-like form have been utilised for HRI. Some researchers create their own robots such as Loki and Muecas [Cid and Nunez, 2014] who were created for the purpose of being able to recognise and express facial emotion while others use commercially available robots including Pepper and Nao from Softbank Robotics, Jibo from MIT [Émond *et al.*, 2020], PR2 from Clearpath Robotics, and PARO the seal from AIST [Zaballa *et al.*, 2021]. However, in the case of commercially available robots, it is important to take their hardware capabilities into account when applying them in HRIs. Not all robots are created equally and thus the limits of their hardware can also affect the quality of the interaction, whether it's if the robot has arms [Émond *et al.*, 2020] or quality of the camera resolution and other sensors. Therefore, in this paper, we will be using multiple robotic platforms such as PR2 (Figure 1) and the Pepper robot (Figure 3).

3 Context Emotion Recognition in HRI Hypothesis

As previously discussed in Section 2, EMOTIC and the idea of using context to help recognise emotions has only been researched in psychology and computer science. In the HRI field, the challenges are different and there is an opportunity to explore additional information. Therefore, there is a novel opportunity to apply it to HRI and social robots. We hypothesise that context emotion recognition will be effective in HRI as it has proven to be in other research fields. Its use with social robots will allow for more accurate emotion recognition during interactions further allowing for the robot to respond more naturally. As there are currently no HRI emotion recognition datasets for the purpose of teaching robots the emotions present within a scenario, EMOTIC provides the closest option with its inclusion of more than the 6 basic emotions and context emotion recognition. Therefore, this research will build on the EMOTIC model in order to integrate context emotion recognition into HRI and work towards our hypothesis.

The EMOTIC model is publicly available to researchers and is a Convolutional Neural Network model consisting of two feature extraction modules. One for the visual-scene context and another that extracts the

body of each person in the frame. A fusion module then outputs the list of recognised emotions and VAD scores. Inference can be performed on both images and videos, and will work for multiple people in the same frame. The original code is written in Lua and has been converted to a PyTorch version by their collaborator [Tandon, 2020] which is the version this project will utilise as a base for its use with social robots.

Utilising the PyTorch model on Google Colab achieves an average precision score of 24.979. This score reflects the substitutions made during the conversion from Lua to Pytorch, the body bias issues, and the uneven distribution of emotions in the dataset demonstrating the complex problem of context emotion recognition. However, due to the fact our focus is HRI and social robotics and not computer science, research towards improving the dataset and model will not be prioritised. Instead additions to the base model necessary for its application into HRI are the focus of this research.

4 Emotion Probabilities

A step towards HRI application involves additions to the models output processing. In the future we aim to have robots understanding emotions while interacting with humans, and thus instead of a single image, the robot will need to be able to use a sequence of images or a video segment. This is due to the fact emotions are not always instantly understood in a single frame. Emotions can last longer than a few seconds and there's a period of time where a transition occurs between emotions. During this transition, emotions can be unrecognisable to even humans until a distinguishable facial expression or gesture is made. Therefore, we investigated modifying the existing emotion category predictions by integrating them into a time-series.

By adding this time-series aspect into the output, the goal is to be able to track the probability of each emotion over the time of the interaction in order for the robot to recognise the current emotions and later on react to the most prominent one. A graph of emotion probabilities plotted over time is chosen as a visual representation for debugging purposes and for future use in HRI experiments. For example, during an interaction this graph can be used to visually communicate to the person interacting with the robot which emotions are being perceived by the robot. The person can then provide feedback when the robot is not accurately recognising the correct emotions and take less offence if the robots response does not naturally fit into the interaction scenario [Bartneck *et al.*, 2020]. Therefore, this addition is necessary in order to successfully apply context emotion recognition into HRI.

We probabilistically combine the sequence of image-based data for this multi-label multi-class problem as,

$$P(c|z_{1:t}) = \frac{P(c|z_t)P(z_t)P(c|z_{1:t-1})}{P(c)P(z_t|z_{1:t-1})}$$

where, c is the class label, z is the sensor reading and t represents time. The expression is then further simplified for our purpose to a product of the current measurement of emotions and the prior to update the target posterior. We start with an assumed uniform prior probability of $1/m$, with m the number of emotions learnt.

As the time progresses, this model can drift towards unrecoverable probabilities of some states, which was managed through the introduction of constraints. Further, this will facilitate a reasonable time for transitioning from one emotion to another.

5 Incorporation of the Face Module

As humans, we use facial features to understand someone's emotional state. Therefore, it is believed that facial features can be a complementary addition to context and body pose as used in the EMOTIC model. In the human-robot interaction context, this could be further justified as the robots have the ability to move towards people to actively observe faces, if they are far away or obstructed.

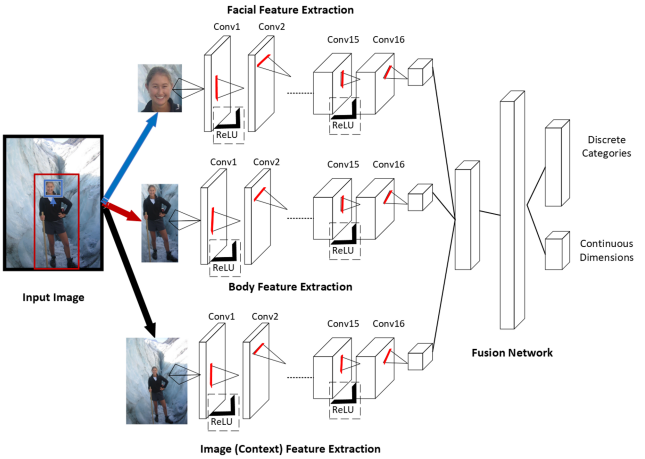

Figure 2: The new HRI-Emotic model with the additional face module.

In order to successfully integrate the EMOTIC base model to include face, body, and context, a face detector was required. The purpose of the face detector is to be used during pre-processing and inference. After investigating a number of face detectors, there were two detectors short listed, Dlib [King, 2002] and MTCNN considering their ease of use and detection speed. With further analysis, it was noted that the Dlib failed to detect a number of the EMOTIC dataset images. This was

most likely due to the nature of EMOTICs images which do contain people far away and with a variety of head angles and occlusions. MTCNN better performed with such images. Further, MTCNN is fast which suits the application [Zhang *et al.*, 2016].

The new end-to-end CNN model is called the HRI-Emotic model and now includes the 3 feature extractors as shown in Figure 2 and is designed to match the proposed model in [Kosti *et al.*, 2017], [Tandon, 2020]. The new facial feature extraction module is modelled after the body feature extractor, except to capture facial features. Similar to the body module, the face module is a Resnet18 pretrained with ImageNet. A custom Loss function and the Adam optimiser is utilised. The fusion module now combines the features of the 3 modules to predict the emotion categories and VAD score, and the parameters of all 4 modules are jointly learnt.

6 Results

In this section, we present the results.

6.1 Evaluation of the Robotic Platforms for the Application

The new HRI-Emotic model was tested with data collected from the PR2 and Pepper robots. We first aim to assess the suitability of the robot's camera picture quality for detecting emotions and secondly, evaluating the test performance.

Data collection was completed in a lab environment with 2 robots. From each robot, we collected images and 30-second videos changing from 'neutral' to 'emotion' to 'neutral' for each of the 6 basic emotions: happiness, sadness, anger, surprise, disgust, and fear. Data was collected with an expert user. The emotions were expressed to the best of the user's ability and were expressed similarly to how they appear in the dataset. The user was standing at a 'comfortable HRI distance' from the robot and in front of a green screen. A green screen is used so that it can be replaced with various other environment types for further study.

The PR2 with its machine-like appearance and face focal point, its potential in interactions, its default cameras and touch sensors, newly added microphone and speakers, and its adjustable height to match most adults, is better suited for the application. Further, its camera has a higher resolution of 2448x2050 at 15fps. Therefore, the PR2 robot, shown in Figure 1, has been chosen as a prospective social robot for testing context emotion recognition in HRI.

The Pepper robot, shown in Figure 3, has also been chosen as a prospective social robot. Reasons include, its humanoid appearance with a face focal point and head LEDs, its potential in interactions, its available tablet and for its default camera, touch sensors and microphone

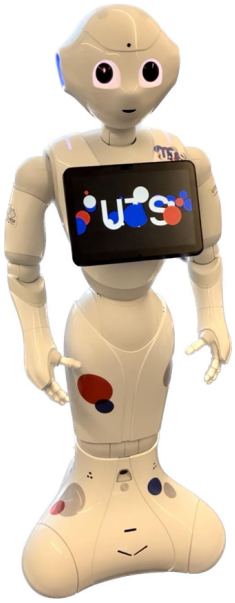

Figure 3: The Pepper robot from Softbank Robotics.

and speaker. Pepper's Top 2D head camera has a resolution of 2560x1080 at 5fps.

Comparing the two results shown in Figure 4, it can be seen that both robots provide a clear RGB picture, although the PR2 provides a crisper image with its higher resolution. This combined with its success in allowing inference to be performed for all the collected data demonstrates each robots suitability for future work with the HRI-Emotic model.

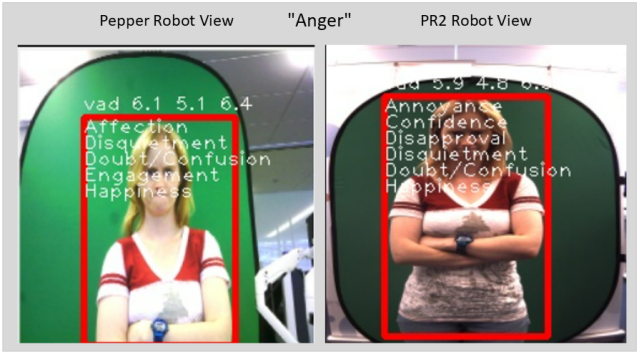

Figure 4: Comparison of Pepper and PR2 collected data for anger.

6.2 Effects of the Face Module

The additional face module in the new HRI-Emotic model has resulted in an improvement over the base PyTorch model's average precision score. Noting that this is a problem with 26 classes, the new model has obtained an average precision score of 25.544 compared to

Figure 5: Comparison of the ground truth labels, the EMOTIC base model and new HRI-Emotic model result.

the base model's score of 24.979, which is a convincing direction towards better emotional recognition. Table 1 compares the results for each emotion class and the total average precision of both models. These results show an improvement in 18 emotion classes, including 5 of the basic emotions: happiness, sadness, disgust, surprise, and fear, by adding the facial feature extractor module.

In the HRI field, granular 26 emotions may not be very useful and we are currently working on identifying a smaller selection of emotions that are HRI task driven to further this research.

Further analysis was completed using dataset images to compare the ground truth labels with the prediction results from the base EMOTIC model and with the new HRI-Emotic model, as shown in Figure 5.

A positive correlation was found between emotions with an improved average precision and emotions being correctly recognised. For example in this Figure 5. Despite the slight improvement, misclassification can still happen due to potential inaccuracies and subjectiveness in the dataset.

Although there is currently no intention to utilise the VAD score predicted by the model, we can also report a slight improvement from the base model's mean absolute error of 96.27 compared with the new HRI-Emotic model obtaining a mean absolute error of 95.96.

6.3 Emotion Probabilities

The purpose of modifying the model output was to start work towards implementation with a social robot. The proposed method of plotting the emotion probabilities over time is intended to provide visual feedback of the most prominently perceived emotions. The performance of this method has been determined by other authors using a similar approach successfully [Berrio *et al.*, 2017], and by the created graphs correctly matching the recognised emotions outputted from the model. As the EMOTIC dataset only contains images, publicly available videos and video data collected from the robot has been used to determine the performance of this method.

Table 1: AP Comparison Between Models

Emotion	Base Model	New Model
Affection	27.14	29.06
Anger	8.8034	8.5951
Annoyance	12.697	14.423
Anticipation	57.148	56.528
Aversion (Disgust)	6.9653	7.0766
Confidence	76.011	75.481
Disapproval	12.42	12.124
Disconnection	22.626	23.437
Disquietment	16.23	15.895
Doubt/Confusion	17.596	17.666
Embarrassment	2.4176	2.1489
Engagement	86.215	86.072
Esteem	15.404	15.493
Excitement	69.042	69.539
Fatigue	9.0144	10.017
Fear	5.8096	5.977
Happiness	66.311	68.214
Pain	6.8397	6.6061
Peace	21.272	22.273
Pleasure	41.426	43.431
Sadness	17.075	18.062
Sensitivity	5.5146	7.0636
Suffering	17.128	19.262
Surprise	8.2521	8.9713
Sympathy	12.163	12.308
Yearning	7.933	8.4297
Total AP	24.979	**25.544**

Figure 6 is a good illustration of the performance of this method, demonstrating how the graph is visualised and how the plotting of emotion probabilities progresses over the course of the input video. The graph shows the most prominently perceived emotion as the highest plotted point. The x-axis is the corresponding time in the video in units of frames and the y-axis is the calculated target posterior, i.e. the normalised fused prior. Next to each image is the emotion list provided by the Emotic model and their corresponding graph colour.

As shown in Figure 6, 4 emotions, out of the 26, have been recognised during the example video by the model. The graphs below present these emotions as the most prominent, and the plotted line switches back and forth between engagement and happiness as the main emotion being perceived. Therefore, it can be seen that the

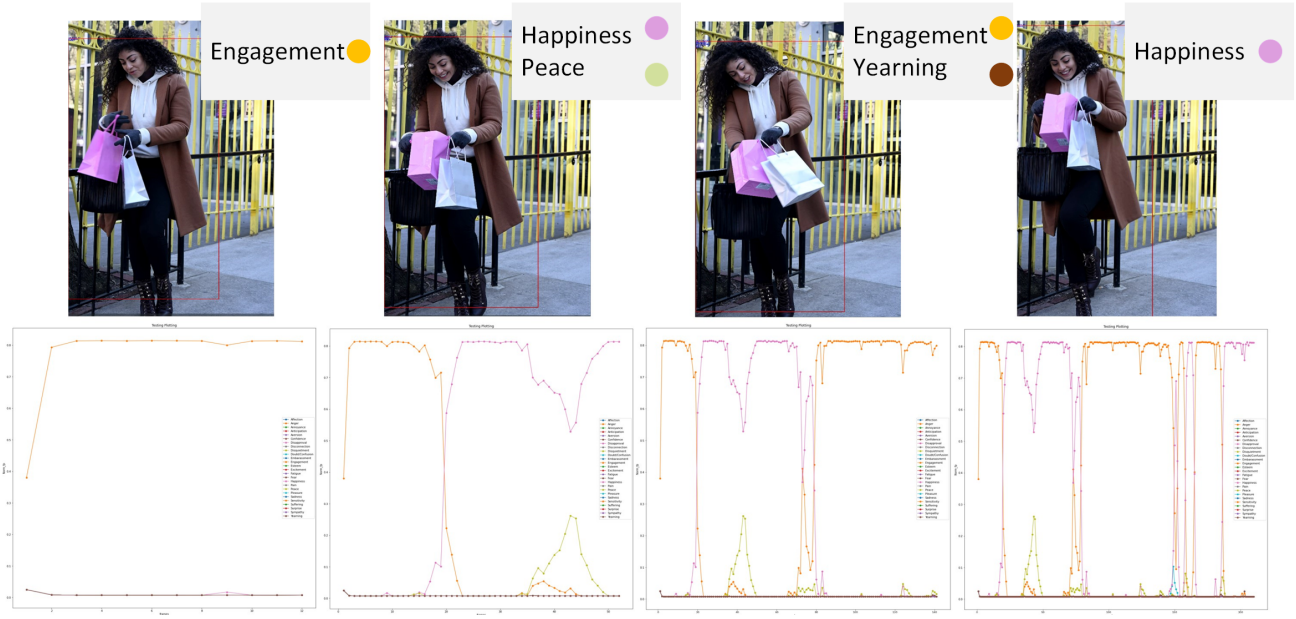

Figure 6: Video stills and the corresponding graphs.

method has successfully visualised the emotion probabilities over time.

For the example video in Figure 6 with the applied probabilistic method, a comparison was also completed to determine the number of emotion labels predicted correctly over the course of the video, i.e. out of all the frames. The average length of the video data used was 10 seconds and each frame was expertly labelled with the most prominent emotion recognised. It was found the base model was 66% accurate while the new HRI-Emotic model was 88% accurate. Thus demonstrating the benefit of the additional probabilistic method for HRI.

7 Discussion

7.1 Overall Limitations of Context Emotion Recognition

The main limitation of context aware emotion recognition, is its low accuracy among EMOTIC and other models. This is attributed to the early stages of the research area, the complexity of the problem, and the lack of extensive datasets. Furthermore, EMOTIC has its own limitations that negatively affect its performance. The dataset contains an uneven distribution of data for each emotion class, with 5 out of 6 of the basic emotions only being present in approximately 1 or 2 percent of people in the dataset [Kosti *et al.*, 2017] resulting in these emotions being difficult to recognise (data imbalance problem).

7.2 Emotion Probabilities

The output processing modification demonstrated in this paper shows a working concept that still has room for improvement. A current limitation of this method is the negative effect it has on the inference time. The creation of graphs in real-time while predicting emotions slows down the inference time considerably and therefore must be fixed before implementation on a social robot in the future. Another possible improvement is the graph aesthetics. The goal is to make it easier to read since we aim to have people use the graphs as feedback during an interaction.

7.3 Incorporation of the Face Module

The new HRI-Emotic model with the additional face module obtained encouraging results even with the challenges in the research field. The results confirm these limitations are being transferred into work towards HRI application. Future work into overcoming these limitations is not planned because our research focuses on HRI. However, in the future changes could be made to the face model including pretraining it with something more suitable than ImageNet in order to improve performance. The large number of classes for the EMOTIC model and dataset may also be a cause for why the model is difficult to improve upon. Therefore, in the future, the number of emotions will be investigated with due consideration given to HRI specific scenarios with social robots. It is expected that the relevant number of emotions to be significantly lower.

7.4 Testing with Robot Collected Data

A number of further research questions arose during the analysis of the data, in addition to the lack of evidence for the second aim. For this paper, data was collected with a green screen blocking out most of the background. This was done because we believed the lab environment wouldn't provide useable information to the model and so further research into comparing the effects of different backgrounds can be done by editing/replacing the green screen with appropriate background images. In the future we plan to move from offline testing to online testing with a robot in order to advance our research in the context of HRI and social robotics.

8 Conclusion

This paper addressed the problem of human emotion recognition in HRI. The state of the art EMOTIC model which was developed by the computer vision community was adapted to HRI while exploiting HRI specific constraints. Rather than using randomly arranged images as done by computer science researchers, a sequence of images (video) were analysed and probabilistically fused to improve the emotion recognition accuracy. The results were promising with the proposed method with 88% accuracy when comparing with 66% accuracy of the base model.

Acknowledgments

This research is supported by an Australian Government Research Training Program Scholarship.

We'd like to thank Muhammad Rihabe for his help in collecting data with the PR2 robot.

References

[Barrett *et al.*, 2011] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, October 2011.

[Bartneck *et al.*, 2020] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Sabanovic Selma. *Human-Robot Interaction: An Introduction*. Cambridge University Press, 5 2020.

[Berrio *et al.*, 2017] Julie Stephany Berrio, James R. Ward, Stewart Worrall, Wei Zhou, and Eduardo Mario Nebot. Fusing lidar and semantic image information in octree maps. 2017.

[Cid and Nunez, 2014] Felipe Cid and P. Nunez. Learning emotional affordances based on affective elements in human-robot interaction scenarios — xv workshop of physical agents, june 2014, leon (spain). 2014.

[Ekman, 1992] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6:169–200, 1992.

[Émond *et al.*, 2020] Catherine Émond, Lundy Lewis, Hajer Chalghoumi, and Muriel Mignerat. A comparison of NAO and jibo in child-robot interaction. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, March 2020.

[Huang *et al.*, 2021] Yibo Huang, Hongqian Wen, Linbo Qing, Rulong Jin, and Leiming Xiao. Emotion recognition based on body and context fusion in the wild. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3602–3610, 2021.

[King, 2002] Davis E. King. GitHub - davisking/dlib: A toolkit for making real world machine learning and data analysis applications in C++ — github.com. https://github.com/davisking/dlib, 2002. [Accessed 10-Aug-2022].

[Kosti *et al.*, 2017] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[Kosti *et al.*, 2019] Ronak Kosti, Jose Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[Lee *et al.*, 2019] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10142–10151, 2019.

[Mittal *et al.*, 2020] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14222–14231, 2020.

[Tandon, 2020] Abishek Tandon. GitHub - Tandon-A/emotic: PyTorch implementation of Emotic CNN methodology to recognize emotions in images using context information. — github.com. https://github.com/Tandon-A/emotic, 2020. [Accessed 02-Mar-2021].

[Yang *et al.*, 2022] Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang. Emotion recognition for multiple context awareness. In *Lecture Notes in Computer Science*, pages 144–162. Springer Nature Switzerland, 2022.

[Zaballa *et al.*, 2021] Karenina Nicoli H. Zaballa, Lance Dean Cameron, and Adrianna Skyler Lugo. Human-robot interactions design for interview process: Needs-affordances-features perspective. In *Interacción*, 2021.

[Zhang *et al.*, 2016] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23:1499–1503, 2016.

[Zhang *et al.*, 2019] Minghui Zhang, Yumeng Liang, and Huadong Ma. Context-aware affective graph reasoning for emotion recognition. *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 151–156, 2019.