Emotion Recognition - A Tool to Improve Meeting Experience for Visually Impaired

Mathieu Lutfallah(✉), Benno Käch, Christian Hirt, and Andreas Kunz

Swiss Federal Institute of Technology, Zurich, Switzerland
{lutfallah,kaech,hirtc,kunz}@iwf.mavt.ethz.ch
https://www.icvr.ethz.ch/index_EN

Abstract. Facial expressions play an important role in human communication since they enrich spoken information and help convey additional sentiments e.g. mood. Among others, they non-verbally express a partner's agreement or disagreement to spoken information. Further, together with the audio signal, humans can even detect nuances of changes in a person's mood. However, facial expressions remain inaccessible to the blind and visually impaired, and also the voice signal alone might not carry enough mood information.

Emotion recognition research mainly focused on detecting one of seven emotion classes. Such emotions are too detailed, and having an overall impression of primary emotional states such as positive, negative, or neutral is more beneficial for the visually impaired person in a lively discussion within a team. Thus, this paper introduces an emotion recognition system that allows a real-time detection of the emotions "agree", "neutral", and "disagree", which are seen as the most important ones during a lively discussion. The proposed system relies on a combination of neural networks that allow extracting emotional states while leveraging the temporal information from videos.

Keywords: Emotion recognition · Neural networks · Non-verbal communication

1 Introduction

Emotion Recognition (ER) is a significant component of non-verbal communication. It allows perceiving another person's reactions, intentions, honest opinion, and mood. Humans recognize emotions by relying on facial expression, voice intonation, and spoken words among other tools. However, blind and visually impaired people cannot perceive the non-verbal cues i.e. the facial expressions [9]. This means that ER's advantages, e.g. to better interpret the reaction and intention of a person they are interacting with [5], are inaccessible for visually impaired people. The following work proposes a solution to this problem focused on in-person and mixed team meetings.

This work was commonly funded by DFG, FWF, and SNF under No. 211500647.

K. Miesenberger et al. (Eds.): ICCHP-AAATE 2022, LNCS 13341, pp. 305–312, 2022.
https://doi.org/10.1007/978-3-031-08648-9_35

Integration of visually impaired people can be improved using Artificial Intelligence (AI) performing ER, which is then conveyed to them. However, ER using AI is challenging. Despite the universality of facial expressions indicating the seven basic human emotions, facial emotional expressions might differ due to the variation in emotional expressivity across cultures and facial features and appearance across ethnicities. In addition, humans are capable of a certain degree to control their facial emotion expression. Previous research mainly focused on ER using video clips taken from movies such as the Acted Facial Expressions In The Wild (AFEW) dataset[1]. This dataset presents clips with distinctively different illuminations and background conditions. The clips are labeled with seven emotions i.e. happy, sad, surprise, angry, fear, neutral, and disgust. However, such emotions are too detailed, and having an impression of primary emotional states such as positive, negative, or neutral is more beneficial for the visually impaired persons in a lively discussion within a team.

In this paper, we focus on video-based facial emotion detection due to two main reasons. First, online meetings could lead to bad audio due to compression or bad microphone, which in turn reduces the visually impaired person's ability to recognize the emotion of the speaker because of missing the upper formant frequencies [17]. Second, it is interesting to detect the emotion of participants who are not speaking and convey that information to the blind and visually impaired participants.

This paper introduces a tool that allows social interaction to be enhanced by making facial expressions accessible to visually impaired people, which is important for communicating sympathy and understanding. After summarizing related work in this field, we will explain our approach in more detail. The remainder of the paper then gives a summary and an outlook on future work.

2 Related Work

To encourage research in ER through AI technology, various groups have created datasets that comprise images, videos, and dialogues with corresponding labels of emotions. In the case of discrete labels, various emotion schemes have been proposed. For example, Ekman [4] defines six universal emotions that are anger, disgust, fear, happiness, sadness, and surprise. However, some datasets such as the EMOTIC [13] dataset, include up to 26 emotion categories.

The EmotiW [2] competition uses the AFEW dataset. The task in this competition is to assign one of seven emotion labels e.g. anger, disgust, fear, happiness, neutrality, sadness, and surprise to each short video clip in the dataset. Unlike other facial expression datasets, the subjects cover a wide age range i.e. 1–70 years, which makes it generic in terms of age. In this competition, state-of-the-art methods were presented for ER and compared based on the test accuracy. The overall accuracy is computed by averaging the accuracy of correct predictions across all classes. Before the rise of the popularity of deep neural networks, frame-level handcraft features were wildly utilized for ER in images.

[1] https://cs.anu.edu.au/few/AFEW.html.

One method proposed to extract a new feature descriptor is called Histogram of Oriented Gradients from Three Orthogonal Planes [1]. They achieved a test accuracy of 45.21% while working with the seven emotion classes. Currently, the neural network based approach generates state-of-the-art performance in all categories of ER from videos to audio and dialogues. The winning method [8] of EmotiW 2017 achieved a 60.03% accuracy. The method consisted of evaluating four networks to extract features from images and a classifier based on audio.

Another dataset is CK+ [11] which consists of the facial expressions of 210 adults, which were recorded using two synchronized Panasonic AG-7500 cameras. The individuals were asked to express a specific emotion which was then labeled regarding the expressed emotion. The clips consist of a couple of frames that start from a neutral facial expression and peak in the frame of the expressed emotion. Samples of this dataset are shown in Fig. 1. These images are close to what can be expected in a meeting since they show a regular background with good lighting conditions.

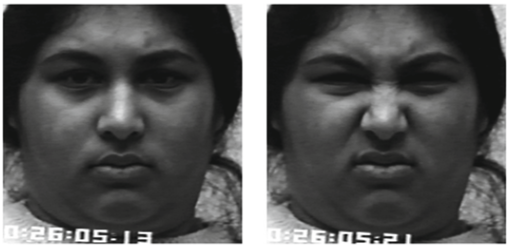

Fig. 1. Example frames of the CK+ dataset at the start of a video and at the end [11].

State-of-the-art methods have been developed for ER, however what is still missing is applications for ER. In one experiment by Marinoiu et al. [12], researchers explore how convolutional neural networks (CNN) can detect the action performed by an autistic child as well as the emotion they are expressing. This would help to design robots that are perceptive to emotion and capable of interacting with autistic children, providing them with a better social experience and helping them to improve. Another application of ER is to help IOT systems interact with users [3]. Moreover, researchers [7] have found that using ER can help in learning environments to maintain student interest.

3 Contribution

To help visually impaired people accessing facial expressions in a net-based meeting, a tool was developed that allows detecting facial expressions, deriving emotional states from them, and delivering the results to the participants. For this, we leveraged state of the art neural networks and public datasets and tailored them for our use case. These datasets contain general emotion categories such as the ones mentioned before.

The pipeline of the tool is to record frames from the facial expression of each participant and then feed those frames into the neural network composed of a CNN, a recurrent neural network (RNN) and a multi-layered perceptron (MLP). Before that, the network required training using the public datasets. To utilize these datasets, we need to cluster the categories of videos they present. The individual steps are explained more in detail in the next sections.

3.1 Network Structure

A video stream of the participant in the meeting is fed to the tool. The first step is to sample the video into frames followed by cropping the images around the face of the participant. The dlib library[2] was used to detect the faces. The images are then cropped and sized to 224 × 224 px. Three frames were fed to the network per clip.

State-of-the-art network structures are then used to detect the emotions. The neural network used, shown in Fig. 2, consists of three parts: feature extractor, fusion, and classifier. The *feature extractor* reduces the dimensionality of the image thus taking out redundant information and keeping only useful one. Resnet-18 [6] was used for this purpose since it gave the best results compared to the VGG [15] network. The output consists of three feature vectors corresponding to the three images that were fed into the network. The *fusion* part combines the features from the three images. The feature vectors extracted from these images are passed to a RNN in order to leverage temporal information. In this work, the gated recurrent network (GRU) is used to combine the information. The three feature vectors are fed sequentially to the GRU allowing to merge information across time. As an output, we get three new feature vectors which are averaged. The *classifier* predicts the probability of the sample belonging to each category of emotions. This is done by taking the averaged feature vector and assigning three scalar values for each emotion class. An MLP with three layers is used to determine these numbers, which are then passed to a soft-max function for normalization. The normalized values are interpreted as the probability of the sample being part of the corresponding class.

3.2 Clustering of Classes

During net-based meetings, only few classes of emotions are expected. Participants in a meeting are mainly neutral except for a few frames where another emotion is shown. Thus, a more general classification of participants' facial expressions is required. The classification includes three classes of emotions i.e. neutral, positive and negative. The public datasets with seven categories need to be adapted to these classes to remove the need to create a new dataset. This

[2] http://dlib.net/.

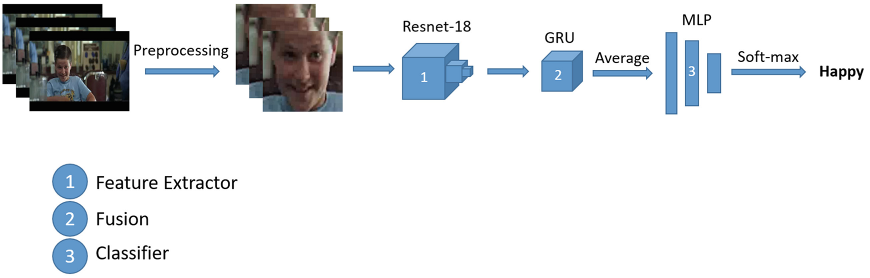

Fig. 2. Pipeline of the tool showing the different types of neural networks used.

can be done by summarizing these categories to have a more general interpretation. As mentioned in Sect. 2, AFEW and CK+ provide samples that are labeled based on seven categories. Thus, we have to group those categories into the three classes we want. The positive group will consist of the happy, surprise categories while the negative one will be formed of the sadness, anger, disgust, and fear categories. The neutral class is already found in the AFEW dataset but that is not the case in CK+ since its replaced by the contempt class. However, as mentioned in Sect. 2, the first frames of the clips are used as neutral frames.

3.3 Network Training

We used the AFEW dataset as pre-training data despite that movie clips present social conditions different to the context of meetings. Pre-training allows having better initial weights for the model thus allowing a better generalization of the model [14]. After pre-training the network using this aforementioned dataset, the network weights were fine-tuned on the CK+ dataset since it shows images of people clearly expressing emotions. Furthermore, these clips were made in more similar environment conditions to our net-based meetings.

One issue we faced is that the network is trained on an imbalanced dataset since the neutral class has much less samples than the negative class as the negative cluster comprises four types of videos which represent sadness, anger, disgust, and fear. Having an unbalanced dataset causes the network to learn only a few of the classes and never predict those with a small number of training samples. To overcome this, we used a weighted loss function, which weights a misclassification on a minority class more strongly than on a majority class. This means that the loss is increased if a neutral sample is misclassified, while the loss is decreased if the negative sample is misclassified.

4 Evaluation and Discussion

The accuracy is computed by feeding clips of videos which were not used for training from the CK+ into the network and then counting the number of correct predictions. The network gave us an 88.8% accuracy on the CK+ dataset and the confusion matrix is presented in Fig. 3. Looking at the confusion matrix, we see that the network correctly predicts the positive emotion with 97% accuracy and the neutral emotion with 98% accuracy. This is in accordance with [10] who has shown that positive expressions are easier to classify since more facial expression actions are involved. The class most challenging to predict accurately is the negative one, in which we only reached 64% while mislabelling as neutral 30% of the time. This result is in accordance with other researchers' work [8]. The high accuracy of the model was expected since having only 3 classes means the random guess lead to 33.3% accuracy versus 14.2% for 7 classes.

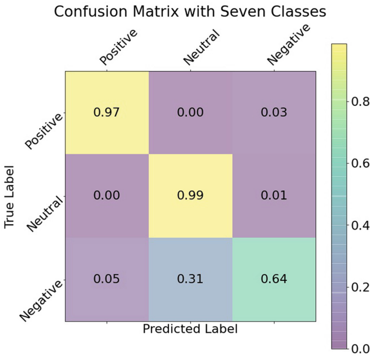

Fig. 3. Confusion matrix for tertiary classification.

The current tool allows displaying the results on a graphic interface as shown in Fig. 4. The scores are shown in terms of a histogram and a smiley face which indicates the emotion. This latter information can later be conveyed to the visually impaired person as a three-stage signal using a Braille display.

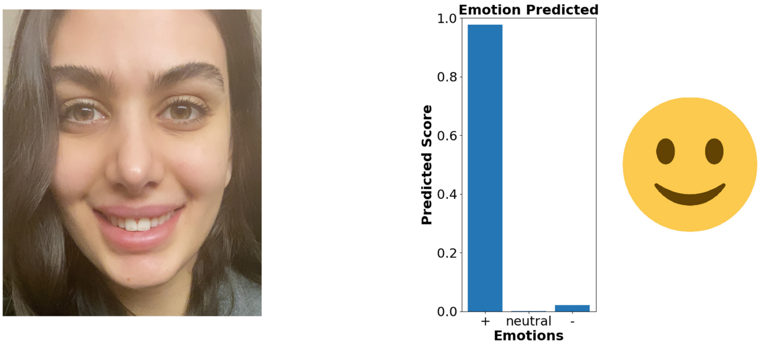

Fig. 4. Happy output of the real-time application.

Interestingly, the network mainly relies on relating specific facial actions to emotions. The network mimic the human ER system, e.g., if teeth are shown or the mouth is shaped in an upward arc, it directly relates to the emotion of happiness. Conversely, the upper part of the face is used to detect sadness since furrowing the eyebrows is usually representative of a negative expression. The absence of any of these two previous features indicate a neutral emotion.

5 Conclusion and Future Work

We introduce a tool that is able to recognize emotional states in a net-based team meeting and by this to enhance visually impaired person's experience in a conversation. This work proves the possibility of using publicly available datasets used for ER to develop such a tool for visually impaired people by clustering the emotion categories to more generic ones. This improvement allows saving time and avoiding the need to develop specific datasets for our use case.

Further investigation can be done on novel network structures and fusion schemes. Resnext [16] is a variation of Resnet that surpasses the latter structure and shows promising results for various tasks. In addition, 3D convolutional networks would be a good extension for our test case since we have data evolving over time. However, these networks still require a lot of memory and large datasets to be functional.

Moreover, a reliable method to output the emotional states to the visually impaired person needs to be investigated. Since the output of recognized emotional states is reduced to a three-stage signal, only two stages need to be output (positive and negative), while the neutral state corresponds to "no output". For this approach, a user study is needed to evaluate the efficiency of the tool in terms of improving the visually impaired people experience. However, due to the pandemic situation this was left as future work.

References

1. Chen, J., Chen, Z., Chi, Z., Fu, H.: Emotion recognition in the wild with feature fusion and multiple kernel learning. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 508–513. ICMI 2014, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2663204.2666277
2. Dhall, A., Goecke, R., Joshi, J., Wagner, M., Gedeon, T.: Emotion recognition in the wild challenge (EmotiW) challenge and workshop summary. In: Proceedings of the 15th ACM on International conference on multimodal interaction, pp. 371–372. ICMI 2013, Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2522848.2531749
3. Dzedzickis, A., Kaklauskas, A., Bucinskas, V.: Human emotion recognition: review of sensors and methods. Sensors **20**(3) (2020). https://doi.org/10.3390/s20030592
4. Ekman, P.: An argument for basic emotions. Cogn. Emot. **6**(3–4), 169–200 (1992)

5. El-Gayyar, M., ElYamany, H.F., Gaber, T., Hassanien, A.E.: Social network framework for deaf and blind people based on cloud computing. In: 2013 Federated Conference on Computer Science and Information Systems, pp. 1313–1319. IEEE (2013)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015). http://arxiv.org/abs/1512.03385
7. Kaklauskas, A., et al.: Affective tutoring system for built environment management. Comput. Educ. **82**, 202–216 (2015). https://doi.org/10.1016/j.compedu.2014.11.016, https://www.sciencedirect.com/science/article/pii/S0360131514002693
8. Knyazev, B., Shvetsov, R., Efremova, N., Kuharenko, A.: Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. CoRR abs/1711.04598 (2017). http://arxiv.org/abs/1711.04598
9. Kunz, A., et al.: Accessibility of brainstorming sessions for blind people. In: Miesenberger, K., Fels, D., Archambault, D., Peňáz, P., Zagler, W. (eds.) ICCHP 2014. LNCS, vol. 8547, pp. 237–244. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08596-8_38
10. Li, S., et al.: Bi-modality fusion for emotion recognition in the wild. In: 2019 International Conference on Multimodal Interaction, pp. 589–594. ICMI 2019, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3340555.3355719
11. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 94–101 (2010). https://doi.org/10.1109/CVPRW.2010.5543262
12. Marinoiu, E., Zanfir, M., Olaru, V., Sminchisescu, C.: 3D human sensing, action and emotion recognition in robot assisted therapy of children with autism. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2158–2167 (2018)
13. Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: Emoticon: context-aware multimodal emotion recognition using frege's principle. CoRR abs/2003.06692 (2020). https://arxiv.org/abs/2003.06692
14. Peng, A.Y., Koh, Y.S., Riddle, P.J., Pfahringer, B.: Using supervised pretraining to improve generalization of neural networks on binary classification problems. In: ECML/PKDD (2018)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015). http://arxiv.org/abs/1409.1556
16. Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
17. Yildirim, S., et al.: An acoustic study of emotions expressed in speech. In: Eighth International Conference on Spoken Language Processing (2004)