INDIVIDUALIZED EMOTION RECOGNITION THROUGH DUAL- REPRESENTATIONS AND GROUP-ESTABLISHED GROUND TRUTH

Valentina Zhang

Phillips Exeter Academy, Exeter NH, USA

ABSTRACT

While facial expression is a complex and individualized behavior, all facial emotion recognition (FER) systems known to us rely on a single facial representation and are trained on universal data. We conjecture that: (i) different facial representations can provide complementing views of emotions; (ii) when employed collectively in a discussion group setting, they enable accurate FER which is highly desirable in autism care and applications sensitive to errors. In this paper, we first study FER using pixel-based DL vs semantics-based DL in the context of deepfake videos. The study confirms our conjectures. Armed with the findings, we have constructed an adaptive FER system learning from both types of models for dyadic or small interacting groups and further leveraging the synthesized group emotions as the ground truth for individualized FER training. Using a collection of group conversation videos, we demonstrate that FER accuracy and personalization can benefit from such an approach.

KEYWORDS

Emotion recognition, facial representations, adaptive algorithm, training data ground truth .

1. INTRODUCTION

In medical practice, emotion recognition is crucial to accurate clinical decision-making [1]. However, there are several obstacles. Patients' emotional behaviors could oftentimes be affected by an underlying condition such as neurodivergence. On the other hand, physicians could be biased by their own emotions and limited by their cognitive ability.

Deep learning (DL) offers a unique value in this context as a DL-based system does not have an emotional bias and could detect patterns and changes too subtle for human cognition in real-time. Yet, there are many challenges for building an accurate DL-based facial emotion recognition (FER) system. Leveraging a small conversing group setting in autism care, our research investigates the potentials of composing DL-based FER systems with dual-representations and automatically deriving the ground truth for quality training data.

This paper hypothesizes that for comparable accuracy rates, pixel-based DL and semantics-based DL sometimes deliver complementing predictions in FER. In the context of small conversing groups, the two types of DL models could be orchestrated to deliver more accurate group and individual emotion recognition. Our adaptive group emotion recognition system includes three components: individual emotion recognition, adaptive group emotion synthesis, and group vs individual emotion modeling. We aim to understand the relationship between pixel-based DL and semantics-based DL and how they could work together to potentially outperform humans in FER.

David C. Wyld et al. (Eds): AIAPP, NLPML, DMA, CRIS, SEC, CoSIT, SIGL - 2022
pp. 47-57, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120904

The main contributions of this paper are (i) a comparatively study of DL models trained with different facial representations; (ii) an adaptive approach toward accurate individual and group FER leveraging discussion group context; and (iii) a proposal to use group emotion as ground truth labels for FER personalization.

The remainder of the paper is organized as follows: In section 2, a few closely related works are presented. In section 3, we outline the two neural networks we used for this work and how their different performances inspired us. Section 4 describes our system architecture, face detection mechanism and working model for adaptive group emotion recognition in detail. Recognizing the established group emotion as a robust ground truth, section 5 outlines how it could be leveraged to improve emotion recognition for individuals and automatic training data labelling. The section then analyzes the test results of our experiment. Section 6 discusses the threats to the validity of our research. Section 7 gives the conclusion and future work of the paper.

2. RELATED WORKS

In the area of individual FER, this work benefitted from studying the pioneering work such as [3][4][14] and holistically understanding their principles and limitations. In the area of group FER, this work has been informed by a diverse set of existing research settings ranging from a four-person UNO game [10] to public crowds [12]. In the area of comparative research in different facial representations and different DL architectures, this work drew its inspirations from [13]. Last but not least, this work relies on [15][16] for the complete and up-to-date survey of all published research works in FER and group FER. With deep appreciation, in this section, we review these representative papers that are most related to and influenced our work.

Alex Krizhevsky et al [14] provides a first detailed description and analysis of architecture and design variables of DL-based image classification. Ian Goodfellow et al [2] has an early yet insightful discussion on the challenges of facial emotion recognition and outlines some important considerations in designing a successful solution. Octavio Arriaga et al [4] explains one of the first and very successful CNN-based individual FER systems. The system does not rely on facial landmarks. Tatsuya Hayamizu et al. [10] is one of the earliest works in group emotion recognition. It also studies a 4-person group, but it relies on classic statistics-based AI techniques instead of DL. Liwei Wang et al. [13], presents a first systematic approach quantitatively characterizing what representations do deep neural networks, and how similar are the representations learned by two networks with identical architecture but trained from different initializations.

3. ANALYSIS OF TWO TYPES OF FER MODELS

When training a DL model for image analysis tasks, there are two general approaches. One is to use the full images as training input data, which is called in this paper as pixel-based DL or image-based training. The other approach is semantics-based, which extracts semantics from the images and uses extracted features such as facial landmarks as the training data. We choose to comparatively study these two different approaches as the human emotion recognition system operates similarly.

In the human brain, a section called the fusiform face area looks at the whole face holistically, which is similar to how a model trained on full images would function. On the other hand, a part of the human brain called the occipital face area recognizes the eyes, nose, and mouth as individual pieces, which would be very similar to a model trained on facial landmarks. The landmark-based training is generally considered more effective because it helps the neural

networks to focus on the essence of the problem, which is the outline of a face. In facial emotion recognition where the outlines to emotion mapping may not be well defined [2], we assume that image-based training may outperform landmark-based training in some cases. This section explains our experiment for validating our assumption.

Our first model is a standard fully-convolutional neural network composed of 60 convolutional and separable convolutional layers, ReLUs, Batch Normalization, Dropout, Flatten, and Global Average Pooling layers. It is trained with the ADAM optimizer, and achieved a validation accuracy of 60% on the FER-2013[3] dataset.[4]

Our second model is trained on facial landmarks which are extracted by solving the shape prediction problem. In this approach, each face consists of a few shapes which outline the face, eyes, mouth, and nose. Through extracting these shapes, this model will ignore the other features and details of the face, which may have introduced noise into the data of pixel-based training. We use a standard 68-point facial landmark system. There are 6 landmarks for each eye, 9 for the nose, 20 for the lips, and the remaining 27 outline the face.

Landmark-based training uses information about these 68 landmarks on the face to correlate the shape of these landmarks with a certain emotion. Four pieces of information from each landmark is extracted: the x and y coordinates, its distance from the mean of all the points, and its vector angle. Collectively, this representation of the 68 landmarks gives a comprehensive summary of facial features. Our second model is a simple neural network consisting of two hidden layers of 128 neurons using the Rectified Linear Unit activation function, and Adamax optimizer. This produces a validation accuracy of 58%.

For our experiment, we analyzed a well-known deepfake video of Robert Downey Jr. created from a speech by Elon Musk using DeepFaceLab 2.0 Quick96 at 1 million iterations. We apply our two trained models to every frame of the video and produce a probability value for each of our seven core emotions. Subsequently, we correlated the results from both models and constructed correlation heatmaps as shown in Fig. 1.

The correlation analysis reveals some interesting results. On the two heatmaps, the correlation between the original and deepfake videos' corresponding emotions is shown as a diagonal orange line of square cells. Our data shows that the landmark-based training model detected a much higher correlation between the original and the deepfake, even in the "Disgust" and "Surprise" components, which are very minimal in the videos.

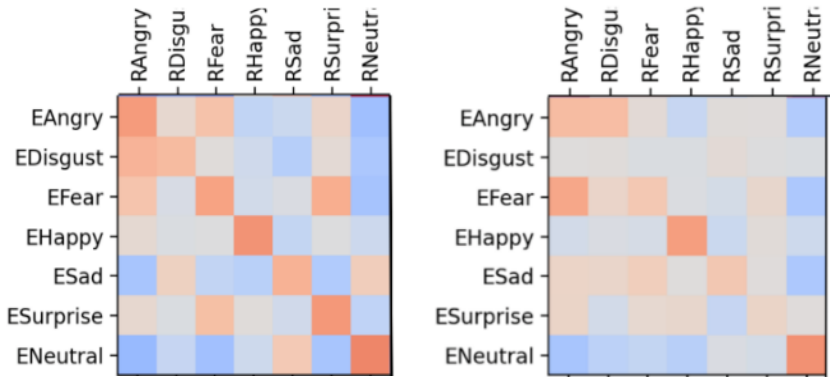

Fig. 1. Emotion Correlation by Image-Trained (left) vs. Landmark-Trained Model (right)

We believe that this is due to the fact that subtle facial expressions are difficult to accurately capture through facial landmarks. On the other hand, the image-based training model detected details that were lost during the creation of the deepfake. To confirm this belief, we employed three human evaluators of the video. Given the manually labeled video from each of the three evaluators, we computed the two-judges agreement to obtain the true labels (e.g., a label was marked as a true positive if at least two of the three evaluators classified it as such). As our assumption predicted, the emotion ground truth on the two faces, as shown in Fig. 2, do not match well. Generally, Elon Musk appeared to have many more positive emotions than the deepfake did.

Our manual analysis also confirmed some subtle or mixed emotions lost in the deepfake video.

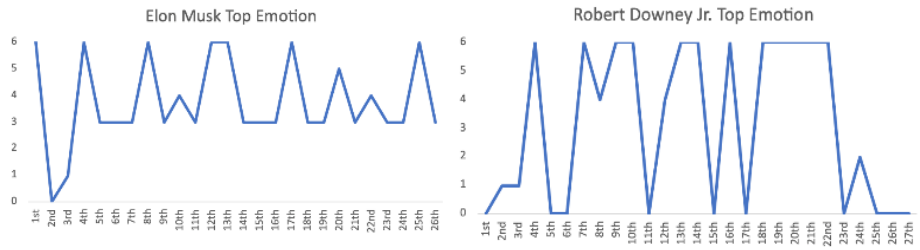

Fig. 2. Ground Truth for Deepfake (left) vs. Original (right)

To analyze our assumption that our two models have complementing strengths in detecting facial emotions, we constructed the complementarity metrics for the recognized emotions as shown in TABLE I.

Tabel I. Dl Models Complementary Metrics

Deepfake	Landmark Correct	Landmark Incorrect
Image Correct	12	20
Image Incorrect	16	4

Exclusive	%
Landmark	42%
Image	33%

Original	Landmark Correct	Landmark Incorrect
Image Correct	8	10
Image Incorrect	6	28

Exclusive	%
Landmark	42%
Image	25%

With the two tables on the left, we report the intersection of sets of true positive emotion detected by two models. For example, 23% of the true positives in the original video and 15% of the true positives in the deepfake are detected by both the image-trained model and landmark-trained model. The relatively small overlap (no more than 51%) suggests that these representations complement each other.

The two tables on the right show the difference in the sets of true positives detected by the two representations. For example, 42% of the true positives detected by the landmark-trained model were not correctly identified by the image-trained model for both the original and the deepfake.
In summary, while the image-trained model picks up more textual details and recognizes more subtle emotions, the landmark-trained model detects less noise and has better accuracy in detecting well-articulated emotions. Thus, there is a potential to create a more effective emotion recognition system by combining the two models.

4. ADAPTIVE FER LEVERAGING GROUP EMOTION CHANGE CADENCE

With its wider adoption, DL-based FER today is used in technologies interacting with humans where accurate detection of individual and group emotion becomes desirable or even necessary [5]. Our motivating interest in better autism care is one example. Group emotion is a complex function of group members' emotions, group context, and environmental context. While the context information is relatively easy to acquire, people reflect their emotions facially in different ways and varying degrees of intensity. It is a challenge to design a system that accounts for individual behaviors.

Taking on the challenge, we construct a system combining the strength of both models from the previous section. Since group emotion is defined by the cadence of individual emotion changes in an engaging environment, our basic idea is to use the cadence to find an optimal trade-off between the two models.

We chose small (4 people) conversing groups as the experiment context of our system for four reasons: First, compared to static images, videos are more redundant for robust emotion recognition. Second, emotional changes within a conversation group tend to be synchronous which simplifies our system design. Third, small groups are easier for manual emotion evaluation and annotation. Lastly, we found a good amount of 2x2 grid view conversation videos for analysis.

The architecture of our system is shown in Fig. 3. There are three important components: noise reduction in emotional change recognition, group emotion synthesis and change alignment, and adaptive weights of two models' outputs.

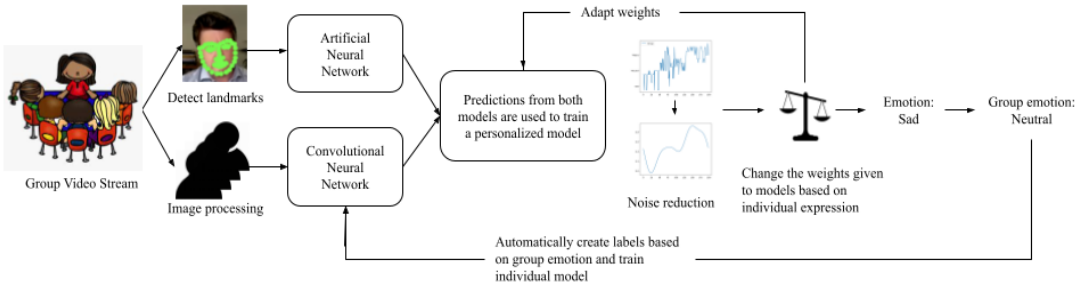

Fig. 3. Architecture of the Adaptive FER System

4.1. Noise Reduction in Emotional Change Recognition

One important limitation observed of the DL-based FER models is that they are sensitive to noisy inputs. For example, image quality jitters from frame to frame in the video stream could cause brief false emotion readings. To combat this, we computed the weighted arithmetic mean of detected emotions for every frame in a second. This averaging allows us to ensure that correlation between changes in emotion are in fact caused by participant's emotions, and reduces the impacts from various input noises.

4.2. Group Emotion Synthesis and Change Alignment

A group's emotion should reflect all its members' emotions. It is more diverse than individual emotions and could consist of more than one dominating emotion reflecting a polarizing group sentiment. To account for that, for any given point in time, our system sums up the probabilities of each emotion type from all group members and uses Euclidean distance to check for deviation. This is adequate for the small group size we analyze. For larger groups, a clustering algorithm such as K-means could be used to separate polarized sections.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

Fig. 4. Euclidean Distance

In detecting emotion changes, we separate the group-wise events where most members change emotions from the individual events where only one member changes. This allows us to treat group-level noises differently from individual-level noises. For group-level noise, we use DFT to convert the time series into frequency-domain and the noise reduction is achieved by eliminating the low-energy frequencies.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i 2\pi k n / N} \qquad k = 0, \ldots, N-1,$$

Fig. 5. Discrete Fourier Transform

For individual-level noises, we want to handle them more carefully because we do not want to miss any real emotion readings obscured by personalized behaviors. We adopted a collection of heuristics such as checking for recurring patterns before dismissing it as a noise.

4.3. Adaptive Weights of Two Models' Output

By our observations, the most common inconsistency in recognized emotions is the intensity in which people express their emotions. One's faint twitch of the cheek may convey the same amount of happiness as another's wide grin. As we demonstrated in earlier discussion, the image-based model is better equipped for recognizing these subtleties which the landmark abstraction could overlook. Thus, if not enough changes in emotions were detected in a certain group member, we increase the weight of the image-trained model to account for more subtle changes. On the other hand, if the emotion readings are noisier than the group average, the group member's facial expressions may be exaggerated, or the image quality may be low. In this case, we amplify the landmark-trained model to focus on the key emotions.

In Fig. 6, we describe our adaptive algorithm. For example, as a baseline, both models are given the same weight. Every minute, we track the amounts of changes in emotion during that minute. Out of the four participants, the ones that were detected to have above-average amounts of changes in emotion were given an extra weight to the landmark-based model and the same weight change is reduced from the image-based model. The opposite was done to the participants who were detected to have below-average amount of changes in the same time frame. Overtime, the weight for each participant settles into equilibrium as the group emotion dynamics converges. To facilitate the convergence, the weight adjustment value is a function of emotion change amount distribution among the group members. The higher the deviation, the higher the adjustment value.

```
Initialize weights for image-based model to 0.5
Initialize weights for landmark based model to 0.5
for every 10 seconds
        for every participant in meeting
                if the amount of emotion changes of member >
                  group number of emotion changes
                        Increase weight for landmark based model
                        Decrease weight for image-based model
                else
                        Decrease weight for landmark based model
                        Increase weight for image-based model
Emotion of each participant = image model prediction * image model weight +
                        landmark model prediction * landmark model weight
Group emotion = average of individual emotions
```

Fig. 6. Pseudo-code for the Adaptive Algorithm

To illustrate the results of our system, the graph to the left in Fig. 7 shows that the adaptive algorithm starts taking corrective action after 20 seconds and has eliminated three transient noises the non-adaptive algorithm classified as "Sad". The graph to the right in Fig. 7 shows that the adaptive algorithm corrected the false "fear" while bringing to surface a subtle angry emotion that would have been missed otherwise.

An issue our adaptive system does not handle well is when an individual's facial structure or neurodivergence makes the person show unintended emotions. In some recordings we reviewed as a part of this work, there were participants consistently misclassified as having a sad or angry emotion component. A more advanced online learning algorithm could aim to detect such persistent patterns in people's emotions and systematically remove the misleading structural component. Alternatively, considering our discussion group context, one could automatically generate personalized DL training data using the group emotion as the ground truth. The next section explores the latter as a solution.

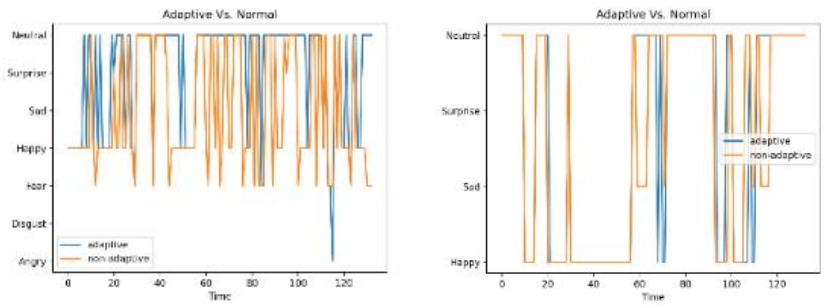

Fig. 7. Emotions Recognized by Adaptive vs Non-adaptive Algorithm

5. GROUP EMOTION AS THE GROUND TRUTH

Group emotion recognition has applications in social psychology [6], shot selection [7], image retrieval [8], surveillance [9], event detection [10], and event summarization [7]. We propose that group emotion be also used for personalizing individual emotion recognition.

For our group emotion recognition purpose, we classify discussion groups into four types: (i) unengaged groups, (ii) engaged groups, (iii) synchronous groups, and (iv) homogenous groups.

An engaged group can establish more information about the emotion correlation among group members than an unengaged group. A synchronous group is a strongly engaged group whose members share the same emotion change cadence, but the specific emotions at a given time may not be the same. At the highest level of engagement, in a homogenous group, group members tend to share the same emotions at any given time.

In the previous section, we discussed how group emotion cadence assists our adaptive algorithm to recognize individual emotion more accurately. It assumes that emotions of the members of the group are all affected by group-wise events, which would indicate that it is a synchronous group. This assumption required us to identify and eliminate the unengaged members from the group emotion calculation.

If we further restrict our application context to a homogenous group, such as in classrooms, or movie theatres, or concert halls, we hypothesize that the group emotion could be treated as the ground truth and used to label new training data. For example, when a group member's images are labelled with this ground truth, DL model could be trained against the member's personal facial expression patterns. In this section, we discuss our experiment designed to validate this idea.

Our experiment consists of the following four steps:

Step 1. We identified group discussion videos of knowledge-sharing nature where group members' emotions are highly synchronized without any divergence of opinions.
Step 2. Using the first half of the video, we compute the group emotion Gx using 3 of the 4 people in the group and label all image frames of the fourth person using Gx as the ground truth.
Step 3. Train our emotion recognition model with the labelled images obtained in Step 2.
Step 4. Using the second half of the video, we apply the newly trained model to the fourth person and check its performance against the models described earlier in this paper.

In principle, Step 3 could be achieved through active learning techniques so that the training could happen online in real-time. However, for this work, we have not tried that because our main goal is to show the validity of using group emotion as the ground truth and the viability of such an approach.

For the validation in Step 4, while a quantitative and general analysis is very difficult, we consider as a qualitative indicator whether automatic labeling increases the correlation between the individual and group emotion. Under our assumption that our group discussion videos do not produce diverging emotions and the group members are evenly engaged throughout the videos, more correlation between the individual and the group would indicate that automatic labeling improves the accuracy of the model. Our correlation metrics are calculated using the Pearson algorithm. For the videos we experimented with, we observed an anecdotal correlation improvement of 15% - 20%. While this is encouraging, we believe more data is needed for a thorough quantitative analysis.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Fig. 8. Pearson Correlation Coefficient

As a specific example, Fig. 9 compares two time series obtained in our experiment. The blue line is generated by the adaptive FER system described in the previous section. The orange line is

generated by the model trained with the automatically labeled data. The main difference between the two is that the blue line was trained with universal data (FER-2013) and the orange line was trained with personal data. Rather than relying on a universal facial emotion model, the orange line is able to identify the inherent "Sad" component in this particular group member's facial expression. Additionally, the false readings of "Fear" and "Angry" due to unrelated facial changes were also detected and compensated.

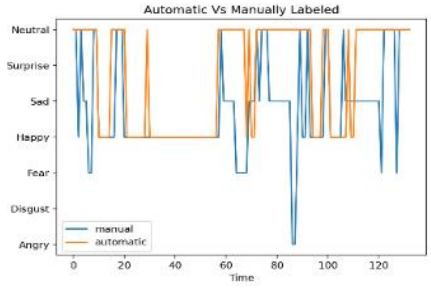

Fig. 9. Model Trained with Auto Labels vs Manual Labels

One important prerequisite for this approach is that the group videos used to collect training data need to exercise the full range of emotion states. This allows the personalized DL model to be trained with labels of all possible emotion values.

Labeling training data is an expensive, error-prone, yet critical step in developing DL-based systems. Our experiment shows an example of how this step could be automated to some extent through the knowledge of group context. Another interesting observation made in our experiment is that data labelled this way captures individual facial emotion patterns which could lead to more accurate and personalized emotion detection. We believe this finding could be applied to other group behavior analysis contexts where DL could play a bigger role in both synthesizing group-level behavior properties and creating individualized DL models.

6. THREATS TO VALIDITY

Construct validity. The main threat is related to how we assess the complementarity of the facial representations: image vs landmark. We support this claim by performing two different analyses: (i) complementarity metrics; and (ii) correlation test.

Internal validity. This is related to possible subjectiveness when evaluating the group emotion in the video fragments used. To mitigate such a threat, we employed three evaluators who independently checked the emotion. Then, we computed two-judges agreement on the evaluated videos. We also qualitatively discuss false positives and borderline cases.

External validity. The results obtained in our study used a small number of selected videos that may not generalize to other small conversing group contexts. To mitigate this threat, we applied our approach to a collection of group discussion videos of different subject areas, such as art, technology, politics, entertainment and sports. Another threat in this category is related to the fact that we apply our approach on pre-recorded videos only. While we do not yet have data to show the effectiveness of our approach in a live group meeting context, the focus of this paper is to show a general technique rather than to build a tool.

7. CONCLUSIONS

In this paper, we show that accurate emotion recognition can be informed by different facial representations. We evaluated the performance of two dominant facial representations and showed their complementary values. Our adaptive group emotion recognition system is flexible and could be reused for different group sizes and contexts. This avoids retraining which eliminates a large time sink native to some DL approaches, and broadens the applicability of our approach. Moreover, as an on-going effort, the adaptive algorithm used by our system is being replaced with an adaptive machine learning (ML) model. Further analysis is being done to assess the relative effectiveness of this ML model against our hand-crafted adaptive algorithm. We hypothesize that the two also present complementary values to some degree, and their complementarity metrics should be studied.

Our approach also highlights the values of accurate group emotion analysis. We showed that by establishing the recognized group emotions as the ground truth, individual emotion patterns such as resulting from neurodivergence could be better analyzed and modeled through automatic training data labeling. This finding speaks to the general possibility of automating the creation of certain training data in various group meeting contexts.

ACKNOWLEDGEMENTS

The author wants to express her heartfelt appreciation to Dr. Neha Keshav and Dr. Ned Sahin at Brain Power LLC for their invaluable guidance throughout this research. The work also received generous support from MIT BeaverWorks Summer Institute. Thank you sincerely.

REFERENCES

[1] Nancy R. Angoff, Making a Place for Emotions in Medicine, 2 Yale Journal of Health Policy L. & Ethics (2002).
[2] Tuan Le Mau, etc. Professional actors demonstrate variability, not stereotypical expressions, when portraying emotional states in photographs, Nature Communications 19 August 2021.
[3] Ian J. Goodfellow, et al. “Challenges in Representation Learning: A report on three machine learning contests.” http://deeplearning.net/icml2013-workshop-competition, July 2013.
[4] Octavio Arriaga et al., Real-time Convolutional Neural Networks for Emotion and Gender Classification, 2016.
[5] J. Bullington. Affective computing and emotion recognition systems: the future of biometric surveillance. Proceedings of the 2nd annual conference on Information security curriculum development. ACM, 95–99. 2015.
[6] P.M. Niedenthal and M. Brauer. 2012. Social functionality of human emotion. Annual review of psychology 63 (2012), 259–285.
[7] A. Dhall et al., 2015. Automatic group happiness intensity analysis. IEEE Transactions on Affective Computing 6, 1 (2015), 13–26.
[8] A. Dhall, A. Asthana, and R. Goecke. 2010. Facial expression based automatic album creation. International Conference on Neural Information Processing. Springer, 485–492.
[9] T. Vandal, D. McDuff, and R. El Kaliouby. 2015. Event detection: Ultra large-scale clustering of facial expressions. In IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Vol. 1. IEEE, 1–8.
[10] Tatsuya Hayamizu, Group Emotion Estimation using Bayesian Network based on Facial Expression and Prosodic Information, 2012 IEEE International Conference on Control System, Computing and Engineering.
[11] James Bergstra and David D. Cox. Hyperparameter optimization and boosting for classifying facial expressions: How good can a “null” model be? Workshop on Challenges in Representation Learning, ICML, 2013.

[12] V. Franzoni, G. Biondi, and A. Milani, "Crowd emotional sounds: spectrogram-based analysis using convolutional neural network." in SAT@ SMC, 2019, pp. 32–36.
[13] Liwei Wang, et al.. Towards Understanding Learning Representations: To What Extent Do Different Neural Networks Learn the Same Representation. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.
[14] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pp. 1097–1105, 2012.
[15] Emmeke A. Veltmeijer et al., Automatic emotion recognition for groups: a review. DOI 10.1109/TAFFC.2021.3065726, IEEE Transactions on Affective Computing.
[16] Wafa Mellouk et al., Facial emotion recognition using deep learning: review and insights. The 2nd International Workshop on the Future of Internet of Everything (FIoE) August 9-12, 2020, Leuven, Belgium.

AUTHOR

Valentina Zhang is a student at Phillips Exeter Academy, an intern at Brain Power LLC developing AI-based technologies for autism care, and an assistant at MIT EEEi workshops building a conductive learning environment for children with neurodivergence. Besides her current responsibilities, Valentina is an alumna of MIT Beaver Works Summer Institute specializing on machine learning in medicine and a speaker at Global AI Student Conference.