Facial Expression Recognition using a Convolutional Neural Network Ensemble

Cecilia G. Morales, Andrew VanOsten, Michelle Zhao
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{cgmorale, avanoste, mzhao2}@andrew.cmu.edu

Monday 14th December, 2020

Abstract

Facial expression recognition (FER) has been extensively studied given its importance in non-verbal communication. It has a wide range of applications such as pain detection in the medical field, drowsiness detection in driver safety, or facial action in the animation industry, among others. FER is a hard problem to solve given that in real world applications people have a variety of colors, skin textures, races, poses, etc. To tackle this problem, we created an algorithm that uses the FER2013 dataset to correctly label facial expressions from seven different emotions: happy, sad, angry, disgust, neutral, surprise, and fear. We trained nine identically-structured convolutional neural networks (CNNs) with modified versions of the training dataset by adding different image transformations to each one, we then trained a wide standalone CNN, and a standalone 4-CNN using augmented data by adding to the training set cropped versions, horizontal flips, oversampling to twice the size and normalization to each image. The models were combined in a final ensemble as a weighted sum of the individual network outputs. Our final test accuracy was 67.7%. This result outperformed a human performance baseline. Implementation of the code is at `https://github.com/mzhao98/emotion_recognition`.

1 Introduction

Human communication consists of much more than verbal or written elements. Non-verbal communication conveys messages by distinct ways including: gestures, body language, facial expressions, and eye contact, among others. Facial expressions play a significant role in interactions and behaviors. They are so important that Darwin in 1872 proposed that facial expressions of emotion may have a link to evolution. Given their importance, many fields such as psychology, artificial intelligence, computer vision, and pattern recognition have been extensively studying them. Facial expression recognition (FER) can have applications in multiple fields such as human-computer interaction, virtual reality, augmented reality, education, and audience analysis in marketing and entertainment. It can also have applications in driver safety such as drowsiness detection, pain analysis in health care, video conferencing, credit card verification, criminal identification, facial action synthesis for the animation industry and cognitive science [4].

Even though facial expression recognition has attracted many scientists, real world applications have a long way to go and are not widely used. These systems have yet to evolve and need to improve their accuracy. Since people in real world applications are from various races, colors, and skin textures, the recognition features may differ [4].

Considering that it is very difficult to create a FER that tackles the aforementioned differences, in this paper we investigate how to detect human emotion from images of their facial expressions using the Facial Expression Recognition 2013 (FER-2013) dataset. The dataset was created by Pierre Luc Carrier and Aaron Courville by using the Google image search API to search for images that matched the six basic emotions—happiness, sadness, anger, surprise, fear, and disgust—which are universally experienced through all cultures according to renowned psychologist, Paul Eckman [1]. The neutral facial expression is also included.

2 Related work

The facial expression recognition dataset (FER-2013) was first introduced as part of a Kaggle contest where 56 teams of competitors were invited to design the best system for recognizing which emotion is best expressed in a photo of a human face [2]. Ian Goodfellow performed some small-scale experiments to estimate human performance on this dataset and found it to be 65 ± 5%. James Bergstra also created a "null"model, consisting of a convolutional neural network with no learning except in the final classifier layer. Using the Tree Parzen Estimator (TPE) hyperparameter optimization algorithm, he found that the best such convolutional network obtains an accuracy of 60%. Using an ensemble of such models, he obtained an accuracy of 65.5%. Out of the 56 teams, four were able to beat the "null" model. The first three teams trained an ensemble of convolutional neural networks (CNNs) with image transformations. The winning solution used a Restricted Boltzmann Machine and the primal objective of an L2-SVM as the loss function for training in order to achieve a 71.162% accuracy. During this contest, there was evidence that CNNs outperform feature-learning algorithms such as SIFT and MKL, but the difference in accuracy was not significant [2].

Yu and Zhang proposed a deep, 7-hidden-layer CNN-based facial expression recognition method.The network contained five convolutional layers, three stochastic pooling layers and three fully connected layers [7]. They discovered that the performance of the algorithm could be further improved if they perturbed the input faces with additional transformations. These transformations made the model more robust to spatially-varying or rotated faces. The training images were also randomly flipped for additional robustness. By adding perturbations to the images, they added unseen training samples that were also be used for training. One of the main differences between their work and those described in [2] is that they didn't perform ensemble voting via uniform averaging but instead utilized a weighted averaging with learned weights. They got second place in the Emotion Recognition in the Wild Challenge (EmotiW) 2015 [7].

Some researchers have also tried to find the most important set of features of a face using a dimensionality reduction technique such as PCA and sparse learning, and try and train a model as such. Zhang et al. investigated two types of features extracted from images: geometric positions and with a set of multi-scale and multi-orientation Gabor wavelet coefficients, both taken from a set of fiducials located in the face [8]. Their model was based on a two-layer perceptron. They performed a nonlinear dimensionality feature reduction with the first layer and the second layer made a statistical decision based on the reduced set of features in the hidden units [8].

There has also been extensive research done by Wan, Yang and Li on the interaction between the dataset, network architecture and training techniques to find the best fine-tuning strategy, trade-off between training speed and complexity of the network, and whether to use a pre-trained model or train one from scratch [6]. It was demonstrated that including different regularization techniques such as dropout and data augmentation suppressed the over-fitting issue. Another finding was that during fine-tuning, freezing more layers resulted in performance degradation compared to fine-tuning weights across all layers. And given the size of the dataset of FER-2013, training a model from scratch took about the same time as fine-tuning a pre-trained model. They also concluded that preserving the size of feature maps at early layers is crucial for images with small size. Their trained CNN reached an accuracy of 65.3% in the test data of FER-2013 [6].

The question raised by Norden and von Reis Marlevi was whether a higher accuracy in facial expression recognition could be achieved if the seven emotional classes were combined into two classes[3]. Using the ResNet architecture, their test accuracy on the FER2013 dataset of 0.7679 confirmed that it is possible to raise accuracy in facial expression recognition by combining classes [3].

3 Data

The dataset consists of 48x48 pixel grayscale images of faces. The faces are approximately centered, and scaled so that each face takes up approximately the same amount of space on the image. There are 28,709 training images, and 7,178 testing images. The class distribution among the emotions is nonuniform, or imbalanced in the sense that there exist a large quantity of Class:HAPPY and Class:SAD faces, while Class:DISGUST contains very few examples in both the training and test sets. The dataset proves to be challenging since features such as the age, pose, or race of the person in the picture vary significantly; thus reflecting more realistic conditions.

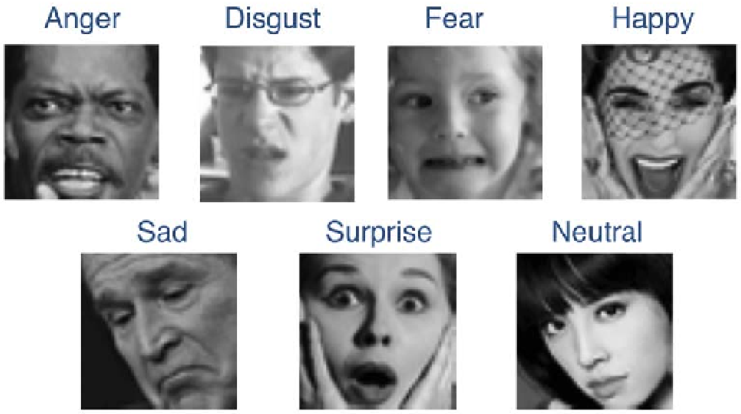


Figure 1: Examples of the seven emotions in the FER-2013 dataset [5]

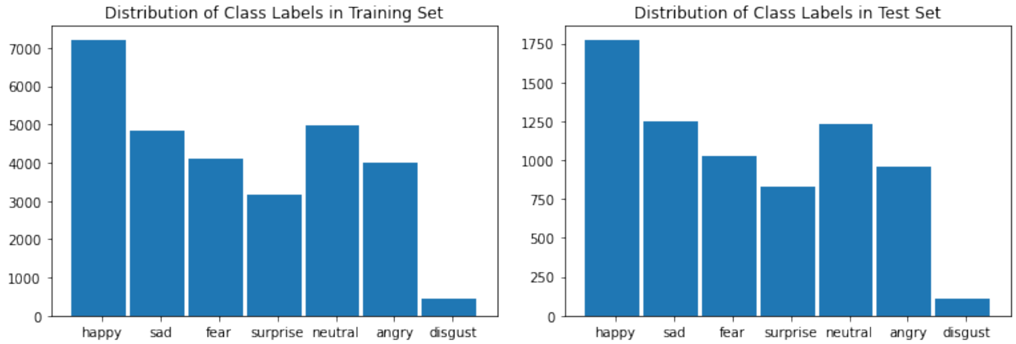


Figure 2: Distribution of class labels in training and testing sets

3.1 Data Augmentation

By performing data augmentation, we increased the diversity of the dataset by applying random transformations and perturbations. By creating a larger, more difficult dataset, we were able to train more robust models. The transformations that we applied were flips, rotations, brightness and contrast adjustments, rescaling (oversampling), cropping, and intensity normalization.

4 Methods

4.1 CNN Ensemble with Image Transformations

For this approach, nine identically-structured convolutional neural networks were trained on nine modified versions of the training data set. Brightness, contrast, noise level, blur, and rotation were adjusted equally across all images to create the following derived datasets: dark, bright, low contrast, high contrast, high noise, Gaussian blurred, right tilted, and left tilted. The unmodified set of images was also used. This approach was chosen in an attempt to create an ensemble of models which are each tailored to different types of input images: one network should be well suited to bright images, another to noisy images, etc. The structure of each network is shown in figure 3. Each weight layer was followed by a batch normalization layer and a rectified linear unit (ReLU) activation function. The first and second linear layers were followed by dropout layers with an associated probability of 0.2.

Each network was trained for 36 epochs with an initial learning rate of 0.001, which was decreased by a factor of 10 at 12-epoch increments. At test time, the normalized network outputs are combined by summing all nine output vectors. The final prediction of the ensemble is the class corresponding

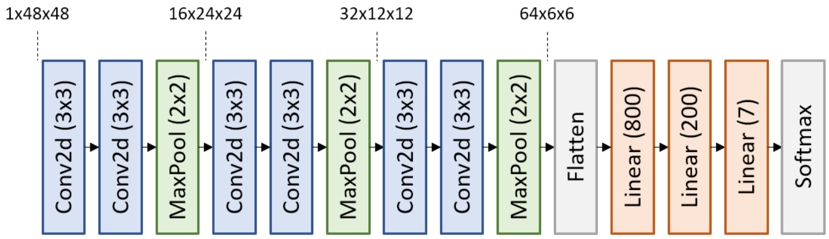


Figure 3: Structure of CNN for image transformation ensemble

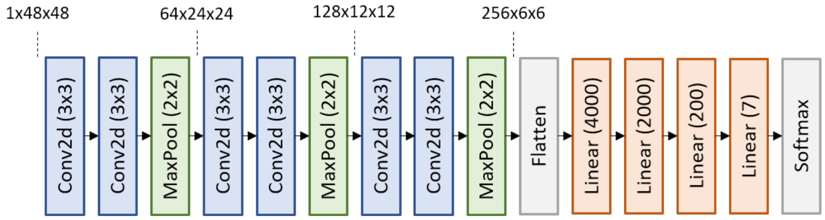


Figure 4: Wide standalone CNN structure

to the largest element of this vector. This method captures not just the predictions of the networks, but also the "confidence" encoded in the full output vector of each network.

4.2 Wide Standalone CNN

With the goal of creating a non-ensemble-based model to compare to the method described in the preceding section, a more complex, standalone CNN was developed. The structure of this network is shown in figure 4. Once again, all weight layers were followed by a batch normalization layer and a ReLU activation function. For regularization, the first three linear layers were additionally followed by dropout layers with probabilities of 0.65, 0.65, and 0.5 respectively.

In an effort to maximize generality of the model, the following random adjustments were applied to augment the training data: horizontal flip, brightness shift, contrast shift, and rotation ($\pm$15 degrees).

4.3 Multi-step Convolution Neural Network

Due to the uneven distribution of classes in the dataset, there exist an extremely large number of examples in the HAPPY class, but a tiny number of examples in the DISGUST class. We observed that the specific emotion classes can be further categorized into 3 groups: Positive, Neutral, and Negative emotions. The Positive category contains HAPPY and SURPRISE. The Neutral category contains only NEUTRAL. And the Negative category contains SAD, FEAR, ANGRY, and DISGUST. We hypothesized that while human participants, and neural networks alike, might struggle to correctly classify all variants of negative emotion, it may instead be easier to distinguish correctly between only 3 classes of positive, neutral, and negative emotions. Next, once in the negative category, it may be easier to distinguish more specific negative emotions, such as fear and sadness for example.

Thus, we propose a 2-step network. In the first step, the model uses a convolution network to predict whether the face belongs in the positive, negative, or neutral class. Then, in the next step, a network trained to differentiate positive emotions will further decompose the positive emotion faces into either HAPPY or SURPRISED. Similarly, a network trained to differentiate negative emotions will further decompose the negative emotion faces into specific classes. The negative model was trained for 100 epochs, using SGD optimizer, and a learning rate of 0.001. The positive and initial step models were trained for 20 epochs, since they achieved convergence, using the same parameters. Our results found that this multi-step model performed decently on the 7-class overall classification task.

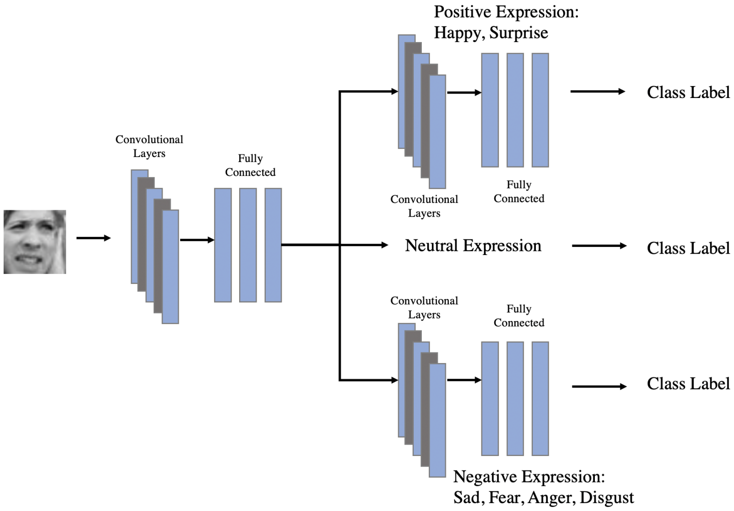


Figure 5: Multi-step CNN structure

4.4 Model 4: Standalone 4-Convolutional Layer CNN

We created a standalone CNN trained using augmented training data. We hypothesized that a simple CNN with moderate width and depth would be able to perform decently on the task by leveraging heavily perturbed data. The model contained 4 convolutional layers of 3x3 kernel size, and max-pooling. The convolutional layers were followed by 3 fully connected layers with ReLU activation and 0.1 dropout. Weight normalization was performed following the first two convolutional layers. In order to train the standalone 4-Conv CNN, the training dataset was doubled by applying a random crop, horizontal flip, oversampling to 2x the size, and normalization to each image, thus creating a perturbed duplicate dataset. The model was trained for 100 epochs, using SGD optimizer, and a learning rate of 0.001.

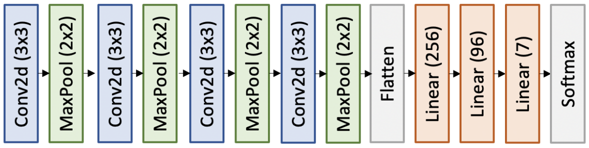


Figure 6: Standalone 4-convolutional layer CNN structure

4.5 Weighted Voting Ensemble

With the goal of getting the best possible performance from these models, one final ensemble was created to incorporate all of them, with the exception of the multi-step CNN. The multi-step model was excluded because its structure and operation are entirely different from the other models. Thus, this ensemble includes all nine of the image transformation CNNs, the wide CNN, and the 4-layer CNN. All of these models have normalized (softmax) outputs, so the ensemble output was computed as a weighted sum of the individual network output vectors. The final prediction was taken as the class corresponding to the largest element of this output vector. Although any arbitrary weighting scheme could be applied, we decided on the following: the nine image transformation CNNs and the 4-layer CNN were given equal weight (each assigned a weight of 1), and the comparatively more complex wide CNN was given four-times that weight (a value of 4).

5 Results

5.1 CNN Ensemble with Image Transformations

The final training and testing accuracy values for each of the models in the image transformation ensemble are shown in table 1. Note: the individual model testing accuracy values shown in the table

Table 1: CNN ensemble with image transformations—results by network

Model (training set used)	**Test Accuracy** (on unmodified test set)
Normal	0.62998
Dark	0.56604
Bright	0.62204
Low contrast	0.57175
High contrast	0.60685
High noise	0.62733
Gaussian blur	0.62162
Left tilted	0.54723
Right tilted	0.56255
Full ensemble	0.65826

were computed by evaluating each individual network on an unmodified test set. The final row of the table shows the accuracy of the full ensemble.

5.2 Multi-step Convolution Neural Network

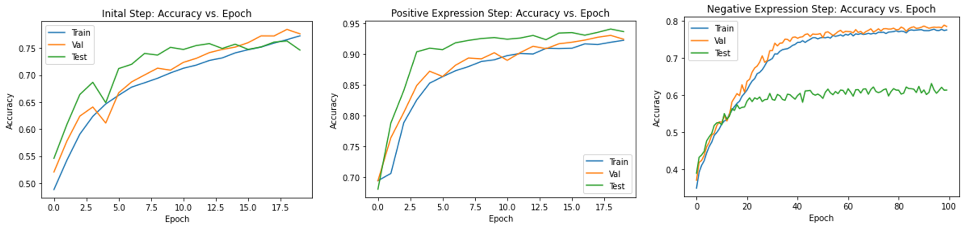


Figure 7: Accuracy vs epoch for positive and negative emotions

We use the term "initial-step" model to refer to the network that makes the coarse classification into positive, neutral, and negative emotion categories. We use "positive-step" model to refer to the network that splits positive emotions into HAPPY and SURPRISED classes. Similarly, we use "negative-step" model to refer to the network that splits negative emotions into SAD, FEAR, ANGER, and DISGUST classes. The initial-step and positive-step models converged after about 10 epochs, so were thus trained for 20 epochs. The negative-step model, which has the most difficult task, converged in about 20 epochs, so were thus trained for 100 epochs.

This model achieved 61.58% test accuracy. The result of note from this experiment was that the multi-step CNN found it much easier to differentiate positive emotions over negative emotions. This result is confounded with the fact that the class distribution meant that the positive-step network simply had to perform binary classification, while the negative-step network had to perform 4-class classification. Still, the positive-network far outperformed the negative-step network.

5.3 Model 4: Standalone 4-Convolutional Layer CNN

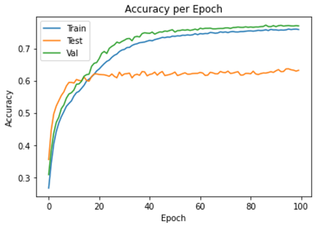


Figure 8: Accuracy vs epoch for standalone 4-convolutional layer CNN

The standalone 4-convolutional layer CNN achieved 63.67% test accuracy, and 99.66% accuracy on the training set. The results of this model exhibit a poor out-of-sample generalizability of CNNs for the facial emotion recognition task.

5.4 Weighted Voting Ensemble

The weighted voting ensemble proved to be the best of our models with a test accuracy of **67.707%**. The confusion matrix for this model is shown in figure 9. This provides a graphical representation of the model's performance across classes. The value in a particular cell represents the fraction of examples in the test set with the corresponding true label (denoted by the row) which were classified as a particular label (denoted by the column). For example, 6% of images with the label "disgust" were misclassified as "fear".

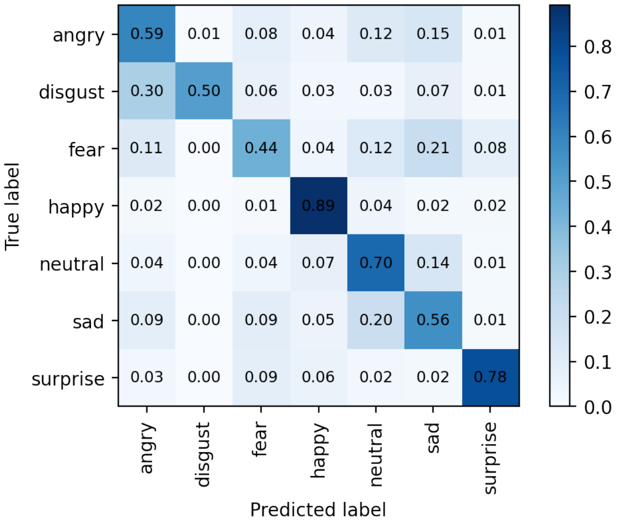
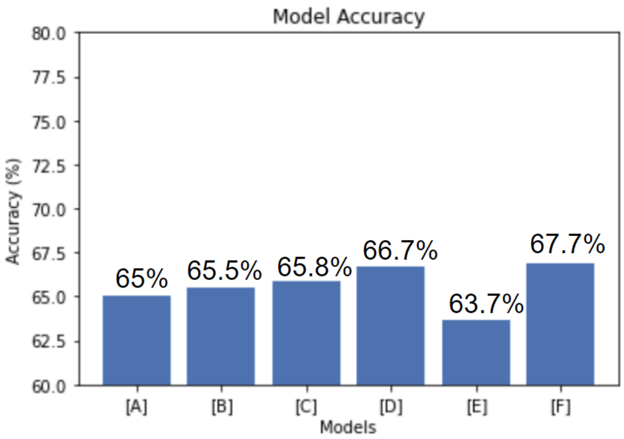


Figure 9: Confusion matrix for the weighted voting ensemble

A summary of the results of our models is shown in figure 10. The 65% human accuracy rate was found by Ian Goodfellow, with an error of $\pm 5\%$ [2]. The null model, developed by James Bergstra, was an ensemble where each model "[consisted] of a convolutional network with no learning except in the final classifier layer" [2]. The best of those individual networks had a test accuracy of 60%, with the full ensemble reaching 65.5%.



[A]	Human Performance
[B]	Null Model
[C]	CNN Ensemble with Image Transformations
[D]	Wide Standalone CNN
[E]	4-Convolutional Layer CNN
[F]	Weighted Voting Ensemble

Figure 10: Summary of results

6 Discussion

As shown in figure 10, several of our models—particularly the ensembles—were able to beat the human performance and null model baselines. Although the multi-step CNN came in just below the baselines, its performance was still quite good. While the wide standalone CNN performed well for being a non-ensemble model, it did not stand up to the best model: the weighted voting ensemble. With a test accuracy of 67.707%, our model would have taken fourth place in the original FER2013 Kaggle competition referenced in [2], trailing the leader by only 3.5 percentage points.

These results provide a good example of the power of ensemble-based models. The nine CNNs that made up the image transformation ensemble had an average standalone testing accuracy of only 59.504%, with the best individual network (the model trained on the unmodified training set) producing a test accuracy of 62.998%. As an ensemble however, these models combined to produce a final test accuracy of 65.826%. Similarly, the weighted voting ensemble beat its best-performing component (the wide CNN) by nearly one percentage point.

The confusion matrix for the weighted voting ensemble (figure 9) shows that, while the model makes plenty of mistakes, many of those misclassifications are quite understandable to the human observer. Some of the most often confused pairs include disgust mistaken as anger, fear mistaken as sadness, and sadness mistaken as neutrality. It would not be far-fetched for a human to make these mistakes when attempting to interpret emotion without further context. Meanwhile surprise and happiness, emotions that correspond to distinctive and often-exaggerated expressions, are rarely misclassified by the model.

7 Future work

Currently, there is not an accurate human benchmark for the classification of the images in FER2013 but instead an approximation calculated by Ian Goodfellow. Thus, it would be interesting to compare the human benchmark with the algorithms. Another future step is to augment the classes with few examples, such as DISGUST, with images found online in order to make learned models more robust and well-performing over all classes.

8 Conclusion

In this work, we develop several learning-based approaches to tackle the task of facial emotion recognition. Facial emotion recognition remains a difficult task. There exists a wide spectrum of emotion representations that are dependent on the human subject's inclinations and face. Even for the task of classifying faces to 7 emotion classes, humans can only achieve around 65% accuracy. Our final ensemble model achieves higher accuracy than the human model, and outperforms all but 3 teams in the original Kaggle competition.

We leverage the following strategies to build a more accurate and robust model: data augmentation, convolutional neural networks of varying width and depth, multistep decision making processes, and weighted ensembling. Ensembling the various approaches helped significantly to improve the out-of-sample generalizability of the overall model on the test set.

Broader Impact

The problem of facial expression recognition through machine learning brings forth the idea of machine Theory of Mind. Such exploration in endowing machines with the ability to understand the emotional and cognitive states of people through computer vision and other techniques offers a breadth of assistance that can be offered to people. For example, for individuals recognized through facial emotion recognition to be sad, automatic support is a potential service that could prove highly beneficial. However, this also raises ethical questions regarding the extent to which machines should be able to understand humans. Through both consideration and development, these questions can be answered, and such services might one day be provided.

We do not expect that our proposed method should leverage discrimination biases in the predictions of the labels since the dataset used is very diverse; however, there were no further studies to ensure that members of different races, ages, or skin colors were classified correctly. Further studies should be done before they are used in real world applications since failure to recognize a particular person could have severe consequences depending on the application, (for example, an old person not having their pain classified correctly in a medical study). There is only one bias that we might expect to see: there is evidence that positive emotions can be more easily recognized than negative emotions. Thus, one should have a calibrated level of trust with the algorithm. We propose to add more negative emotions in the training set to try and offset the bias. Given that our model theoretically outperforms humans, it gives hope that one day computers might be able to take on important tasks such as criminal identification.

References

[1] Paul Ekman. Facial expressions. In *Handbook of cognition and emotion*, pages 226–232. New York, 1999.

[2] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59 – 63, 2015. Special Issue on "Deep Learning of Representations".

[3] Frans Nordén and Filip von Reis Marlevi. A comparative analysis of machine learning algorithms in binary facial expression recognition, 2019.

[4] E.Ramakalaivani S.Shaul Hammed, Dr.A.Sabanayagam. A review on facial expression recognition systems. *Journal of Critical Reviews*, 7:903 – 905, 2020.

[5] Elizabeth Tran, Michael B. Mayhew, Hyojin Kim, P. Karande, and A. D. Kaplan. Facial expression recognition using a large out-of-context dataset. *2018 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 52–59, 2018.

[6] Yang Li Weier Wan, Chenjie Yang. Facial expression recognition using convolutional neural network: A case study of the relationship between dataset characteristics and network performance. 2016.

[7] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, page 435–442, New York, NY, USA, 2015. Association for Computing Machinery.

[8] Zhengyou Zhang, Michael Lyons, Michael Schuster, and Shigeru Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 454–459, 1998.