Facial expression recognition(FER)
Final Deliverable

University of Toronto
Faculty of Applied Science and Engineering
APS360 - Artificial Intelligence Fundamentals
Due: Apr 13, 2022, 5 PM

APS360W-2022 - Team 55
Jack (Yang) Chen - 1005747649
John (Litao) Zhou - 1006013092
Peter(Jiping) Li - 1005983269

Word Count: 2486

Introduction

It is proven during communication that non-verbal communication accounts for up to 55%, while the proportion of word communication is only 7%[1]. As an integral part of non-verbal communication, facial expressions could reflect an individual's emotional state[2]. With the development of family robots, there is an increasing need for robots to understand human moods and provide more intelligent responses[3]. Therefore, to provide users of family robots with a better HCI (Human-Computer Interaction) experience, we intend to develop a Facial Expression Recognition System (hereinafter referred to as FER), that helps recognize human faces and their expressions(Figure 1). As a system involving the search for patterns in large amounts of data of human faces and making classification, this coincides with the mechanism of machine learning and makes it a great approach.

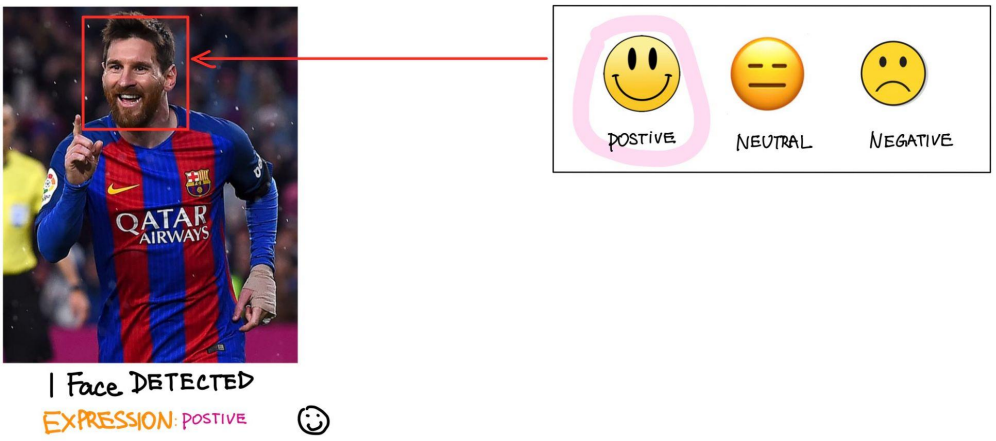

Figure 1:Example of facial recognition and expression classification

Illustration

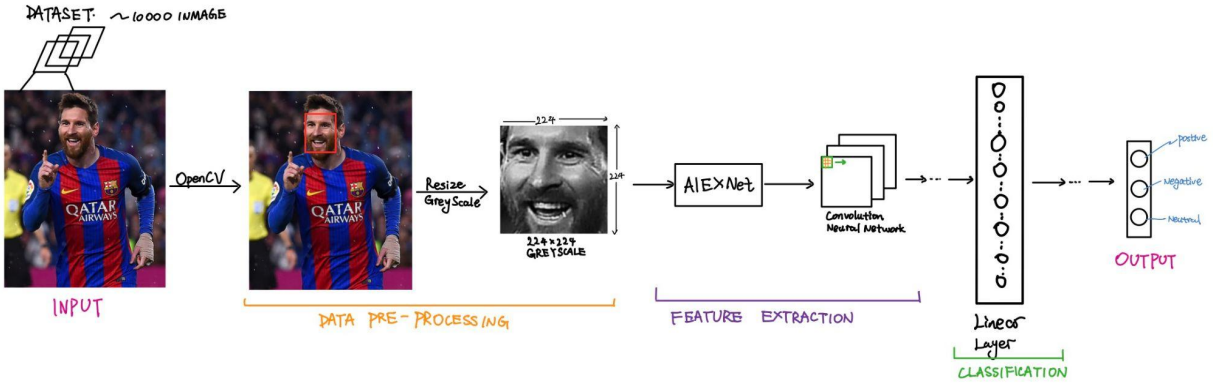

Figure 2: High level illustration of the model

Background

For Facial Expression Recognition, multiple datasets are used as a benchmark. One of the most widely used is FER2013[4], which contains over 30000 preprocessed images of facial expressions in seven categories. Currently, the highest accuracy achieved on this dataset is 76.82%. This model is the winner of an ICML workshop contest[5]. The model is built based on the convolutional neural networks(hereinafter referred to as CNN), and uses SVM and L2-SVM as the loss functions.

Data Processing

Our dataset will be extracted from "Learning Social Relation Traits from Face Images" [6]. The dataset contains 100067 raw images that are in the format of .jpeg/.png, RGB coloured and of varying sizes. Each image is labeled with 18 expressions ranging from {amazed, angry, annoyed, anxious, astound, awe, boring, crying, disgust, distaste, distressed, expressionless, fierce, fighting, frightened, heartbroken, hostile, mad}. A variety of populations are sampled. Most images contain a face and a body and a background and obstacles. Approximately 5% of the images have multiple faces, and 1% do not have any faces(Figure 3).

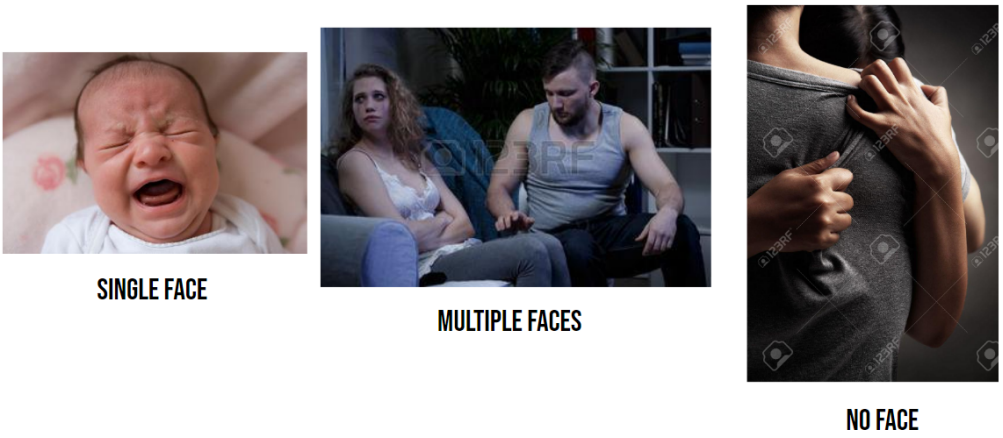

Figure3: Illustration of the dataset

Firstly, we shrink and combine the 18 types of expressions into three emotion classifications types {Positive, Negative, Neutral}. Because through observation, we discover that some face images are incorrectly labeled(Figure 4). As shown, there is no clear boundary between the images from different labels. There are some images with the same labeling that have a huge distinction, there also exist some similarities between images with the different labels.

Figure 4: Example of the similar Images with different labeling

To increase the overall accuracy of our dataset, we conclude that only the "Amazed" and "astound" images can be considered "Positive." Similarly, only the "expressionless" and "boring" images are mapped to the new "Neutral" category. All remaining images from 14 expressions (angry, annoyed, anxious, awe, crying, disgust, distaste, distressed, fierce, fighting, frightened, heartbroken, hostile, mad) are evaluated as "Negative" reactions. The illustration of the structure of our dataset is shown below(Figure 5)

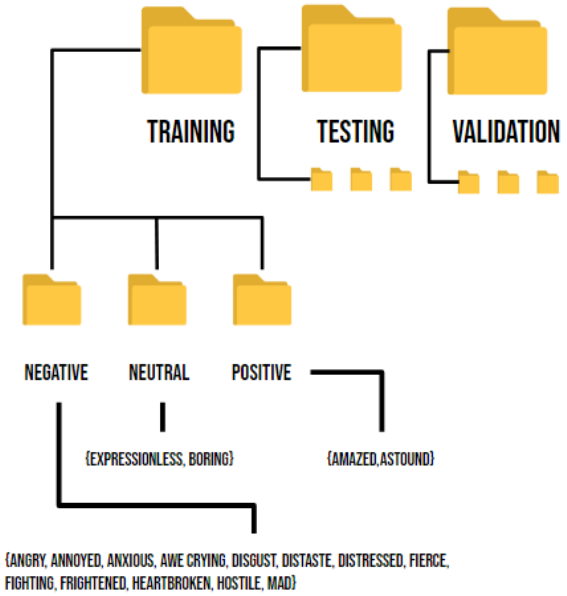

Figure 5: Illustration of the dataset structure

For each image, a grayscale transformation is performed to reduce the color channels from 3 to 1. Then we use a pre-trained Haar-cascade classifier provided by OpenCV as our localization model and determine whether the image contains human face(s); In the case of an image containing multiple faces, we will select the largest one. Then the model will crop the image based on the coordinates

returned by the localization model and then resize it to 224 by 224. The processed dataset is split by a proportion of 8:2:2(training, validation and testing dataset)[7], and the proportion of each expression within each dataset will be maintained. There is no duplication between images in each dataset. The illustration of the entire process is shown below (Figure 6)

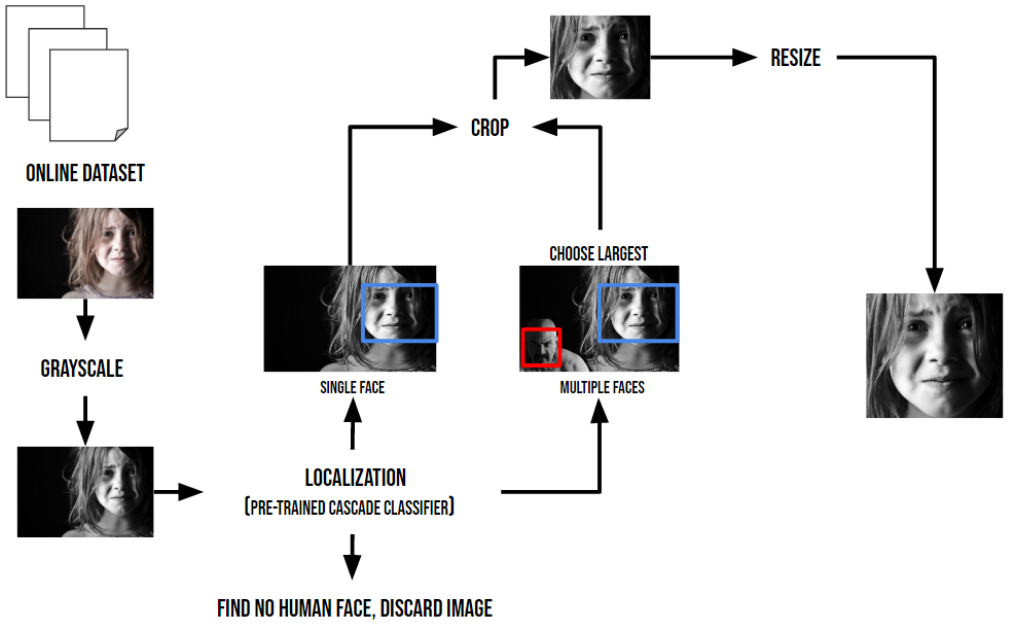

Figure 6: Illustration of the face localization and resize process

Architecture

The first part of the primary model is the AlexNet (Feature transfer learning layer). Each input image will be passed into the AlexNet.features to generate an embedding map of 256x6x6 and they are stored as tensors. Afterwards, these tensors will be passed into two convolutional layers that are designed to be specialized for facial expression recognition.

- The first layer uses 200 3x3 size kernels to extract features from the pre-processed data.
- The last layer uses 150 2x2 size kernels to condense the information. The output of this layer with a size of 150x3x3 will directly connect to fully connected layers.

Afterwards,The classifier is constructed by connecting three fully connected layers with 1300, 200, and 3 neurons(Figure 7). At the output layer, the three outputs (positive, negative, neutral) will use one-hot encoding to choose the largest value as the result.

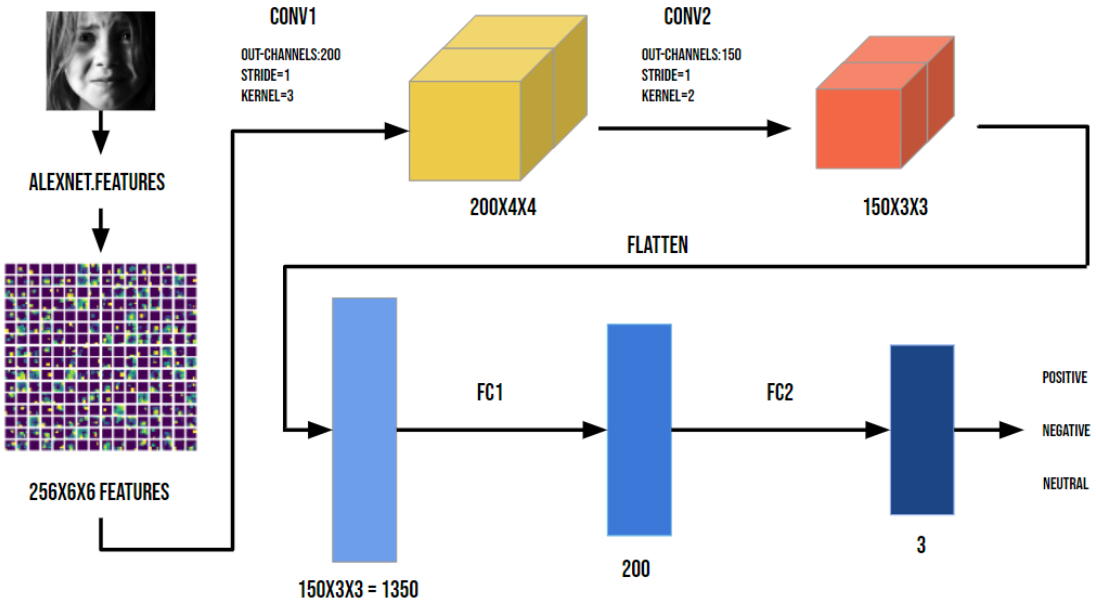

Figure 7: Illustration of the primary model

To introduce non-linearities to our model, the ReLU activation function is applied between each layer. As this is a multi-class problem, we used the cross-entropy loss as our loss function.

Baseline Model

Our baseline model is established based on regular CNN that was used to solve a similar classification problem without AlexNet for the feature extraction. A model consists of 2 convolutional layers, two fully connected layers and one max-pooling layer is established. The ReLu function is selected to be the activation function between each layer. (Figure 8)

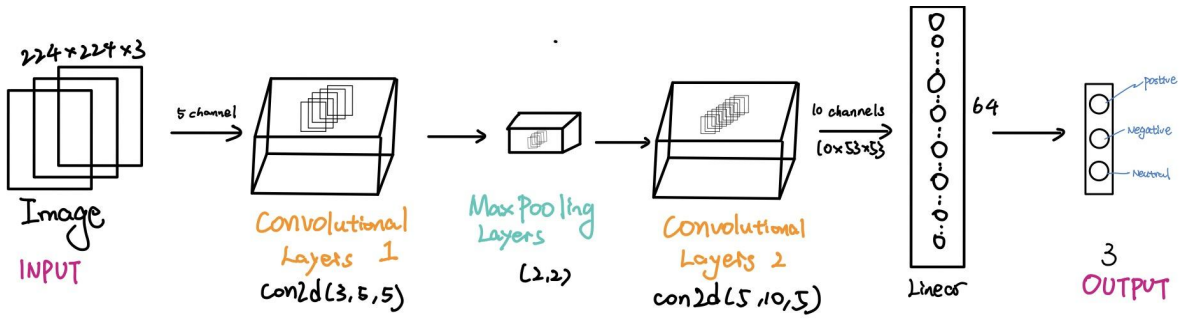

Figure 8: Illustration of the baseline model

The input to the baseline model is the same input we generated from the data processing stage, a fixed size grayscale image. Unlike our primary model, we only applied the two layers of the regular convolutional layer instead of transfer learning using AlexNet for feature extraction. Since the model

consists of fewer network layers, the speed of tuning and training efficiency has dramatically increased.

The hyperparameters used for tuning are as follows:

- Batch Size: 256
- Number of Epochs: 30
- Learning Rate: 0.001

As we load our processed training and validation dataset, a final training and accuracy are reached at 0.5388 and 0.4836, respectively(Figure 9) The accuracy of the testing dataset is achieved at 0.4318.

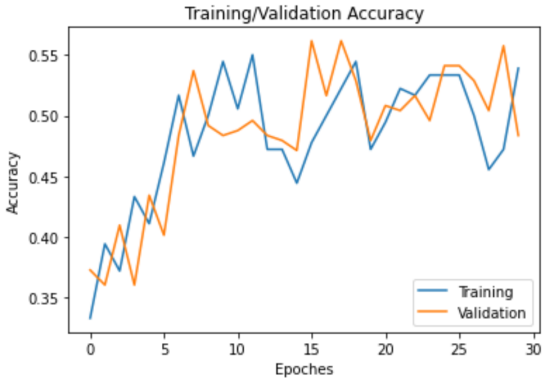

Figure 9: Training and validation curve of the baseline model

These two increasing curves in training, and validation accuracy curves indicate that it is feasible to use convolutional neural networks for this classification problem. Since the input of the baseline model is the same as the input for our own model, the result will be used as the benchmark to evaluate the training quality of our model.

Quantitative Results

The best primary model is trained with the following hyperparameters:

- Batch Size: 256
- Number of Epochs: 60
- Learning Rate: 0.001

Our model is evaluated based on the training and validation accuracy graph and the training loss(Figure 10). Between epochs 0 and 40, both the training and validation accuracy curve increase at a steady rate. Afterwards, training accuracy continues to increase while the validation accuracy curve

fluctuates at around 0.65. The final training and validation accuracy are 0.7458 and 0.6876, respectively. There is no sign of overfitting. On the other hand, training loss fluctuates wildly and then eventually decreases to 0.0013.

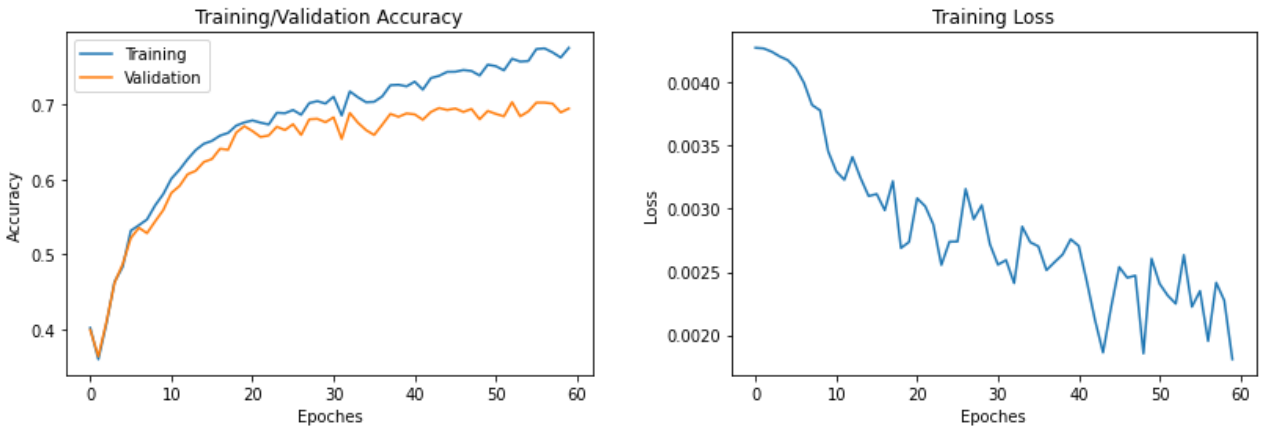

Figure 10: Training and validation curve and training loss of the primary model

Qualitative Results

To evaluate the quality performance of our FER model, we proposed a confusion matrix (Figure 11). It compares the real label and the predicted label of the validation data.

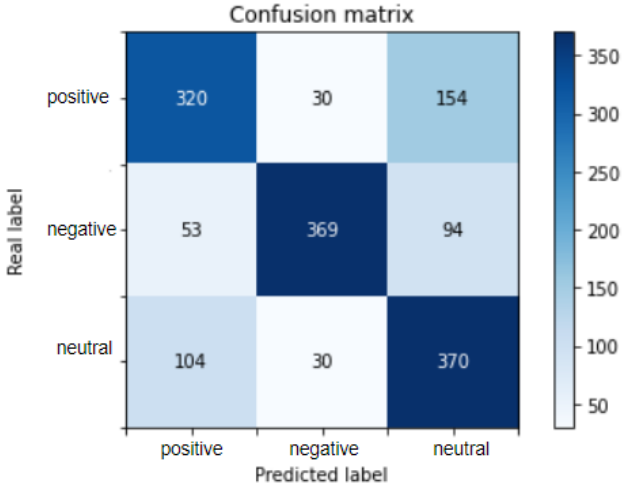

Figure 11: Confusion Matrix for the validation data of the Primary Model

Based on the results, we compute the precision, recall and F1 score for each label, which are defined as follows:

$$\textit{Precision (positive)} = \frac{True\ Positive}{Total\ Predicted\ Positive}$$

$$\textit{Recall (positive)} = \frac{True\ Positive}{Total\ Actual\ Positive}$$

$$\textit{F1 (positive)} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Figure 12: Calculation of the precision recall and F1 score[8]

We compute the precision, recall and F1 score for validation result of our primary model as follows(Table 1) :

	positive	negative	neutral
precision	0.67	0.86	0.59
recall	0.63	0.71	0.73
F1	0.64	0.77	0.65

Table 1 Validation result of primary model

In general, the model is good at identifying negative emotions. However, the model did not perform well in distinguishing neutral and positive emotions. Here are some observations:

- Positive: recall is low, precision is relatively high and F1 is lowest.
- Neutral: precision is low, recall is relatively high and F1 is only 0.01 higher than positive.
- Negative: precision, recall and F1 are all high.

Considering the observed data, we can conclude that the model overestimates "neutral" and underestimates "positive." But positive and neutral results' accuracy is almost the same if we aim for a balance between precision and recall.

Evaluate model on new data

As described in the data processing part, our testing data set is prepared as 20 percent of our entire dataset and is never used in the training and validation stages. For our baseline model, the final testing accuracy reaches 43.18%. We also computed the confusion matrix (Figure 13), precision, recall, and F1 score for the result (Table 2).

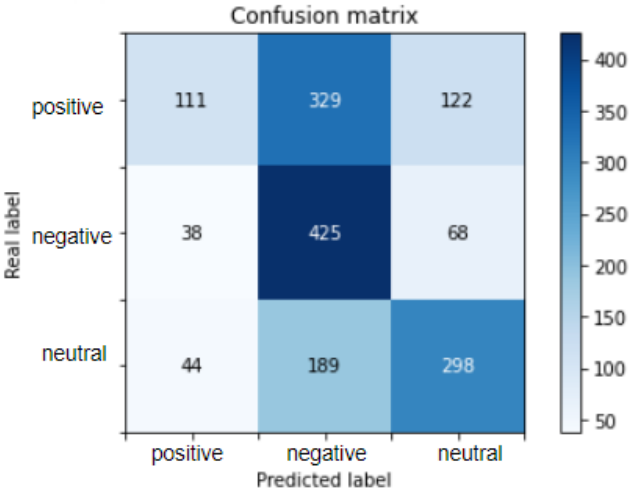

Figure 13: Confusion matrix for the testing data of the baseline model

	positive	negative	neutral
precision	0.57	0.45	0.61
recall	0.19	0.80	0.56
F1	0.29	0.57	0.58

Table 2 Testing result of baseline model

For our primary model, the final testing accuracy reaches 69.21%. In general, the confusion matrix, precision, recall, and F1 score are similar to those obtained from the validation dataset(Figure 14). In comparison to our baseline model(Table 3), we can see that there is a dramatic increase in the F1 for “positive”. In addition, for both “positive” and “negative”, the value difference between precision and recall has narrowed.

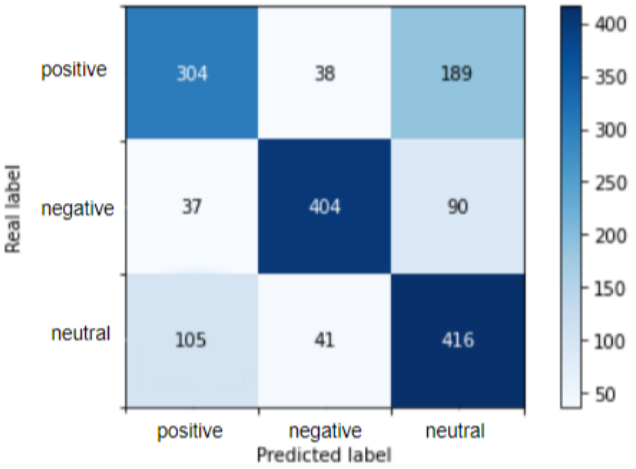

Figure 14: Confusion matrix for the testing data of the primary model

	positive	negative	neutral
precision	0.68	0.83	0.59
recall	0.57	0.76	0.74
F1	0.62	0.79	0.65

Table 3 Testing result of primary model

For our primary model, to evaluate how far the predicted results differ from the real results, we calculate the probability distribution of the predicted results using the softmax function. The table below(Table 4) illustrate the error spread between the incorrectly predicted emotion and the actual emotion.

Error % between predicted and real emotion	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-100%
#Faces	226	219	169	104	37	28	6	1	0

Table 4 Error percentage spread of the predicted and real emotion table

In the table, it's evident that most incorrectly predicted results have an error rate that is less than 20%. It shows that for the model, the boundary between the natural and positive is difficult to discern.

Discussion

Our final testing accuracy achieved 69.21%, which is significantly higher than the testing accuracy of our baseline model. Meanwhile, the accuracy is around 5% lower than the best model based on FER-2013 training data. However, we should consider that FER-2013 possesses a pre-processesd dataset and offers an expression classification of 6 types.

Our experience throughout this project showed the key to creating an effective model is to have a well-labeled dataset. Throughout the experiment, we found that the validation accuracy would only be around 20% if we did not combine 18 expressions into three. We believe that our testing accuracy will be higher if we train our model based on the FER-2013 dataset insead of our own dataset. However, it is not appropriate to use a pre-processed dataset for this project.

Upon reviewing our model, we can make two improvements. We found that several expressions have far more images than others. When we combine 14 expressions into “Negative” emotions. “Negative” will be more closely associated with expressions that have more images. Our solution will either compress those expressions with more images or use GAN or autoencoder to generate new images for expressions with fewer images.

Localization is another area that needs improvement. The pre-trained Haar classifier does not deliver a 100% accuracy rate for our facial recognition model. Around 1% of localized faces actually contain no faces. Several examples are given below(Figure 15).

Figure 15 Example of non-face images that was recognized as faces

Generally, the model sometimes interprets round objects as faces. To solve this, we develop a few ideas to evaluate their feasibility:

- Designing and training our own model. It requires a great deal of facial recognition data and some smart models that fit our problem.
- Picking out the bad images manually during training. It will help the model to learn better. However, we cannot apply this procedure to all testing datasets.

- Add another CNN after the localization model recognizes round objects and eliminates them. We can use a dataset that includes images of watches and tires. However, this CNN could recognize some faces as round objects and eliminate them.

Ethical Considerations

As our topic involves the collection of personal face images, ethical considerations must be taken into account. The consent of participants is required when their image is taken and used for our model [9][10]. This Privacy protection concern should apply to both training data selection and the use of the program. For our dataset, we discovered that all the images we used for training have a watermark on the original image which indicates that the dataset is frequently used for facial expression training.

We need to ensure that our trained program is accurate and unbiased under all circumstances. Since people with different ages, races and genders have different facial features, their data will produce different results after training [9]. As a result, for our data, we try to ensure that these minority groups have enough samples to ensure diversity, and these sample images and labels are unbiased.

Project Difficulty / Quality

The key of our project is to develop a facial expression system that helps recognize human faces and their expressions for family robots as a supplement, providing users with better feedback. Due to the issue of the datasets, our team spends a lot of work working on face locating and cleaning the unusable data from the dataset using OpenCV. We redefine the three categories based on the original and manually divide the data into three classes to improve the accuracy.

Our team establishes a simple CNN as a baseline model, prove the feasibility, and received a testing accuracy of 43% as a benchmark. Based on the baseline model, our team applies transfer learning and used AlexNet to enhance the feature extraction quality. To demonstrate the training quality, our team used the training curve. We also propose a confusion matrix to elaborate on the ability to predict its label and compare it to its real label. Besides, the precision and the recall are also calculated for each category to check the ability of separate different categories.

The final training accuracy is 74%, and the test accuracy is 69% which is slightly lower than the FER 2013 accuracy, while it is significantly better than the accuracy of the baseline model. Based on the result, our model has the highest precision in negative expression with 0.83 compared to other types of expressions. It indicates that the robot equipped with our system is more sensitive to users who

have negative expressions. As our model finds it hard to differentiate between negative and neutral feedback in the application, users are likely to receive negative responses rather than default responses. For the original dataset with 18 types of expressions, more kinds of expressions are combined into negative expressions than neutral expressions. It could increase the rate of success in prediction for the robot application. Thus, our team believes that our system is of great quality and has great difficulty.

Link to our Collab:

We changed the link for our project which is different from the link in the proposal:

Image Resize and Rename:

https://colab.research.google.com/drive/1-77QwnNZFdObfuJH1IvHlyIu2hCuZkuA

Baseline Model:

https://colab.research.google.com/drive/1QnqHYZlT_3EqXRAcyGvCGWY--UuxC5_q

Alexnet Tensor Generation:

https://colab.research.google.com/drive/1qAhQzjq3GoE8YItIWoevrpI-c4dpsGsA

Primary Model:

https://colab.research.google.com/drive/1VCA35MTAdav4hKJJoA0Q3oZllrsPwIuM

Confusion Matrix Generation:

https://colab.research.google.com/drive/1KR48Ls1-HFgLJN9M-qdjPhAYo0j41KW-

Reference

[1] “How much of communication is nonverbal?: UT permian basin online,” UTPB, 03-Nov-2020. [Online]. Available: https://online.utpb.edu/about-us/articles/communication/how-much-of-communication-is-nonverbal/. [Accessed: 09-Feb-2022].

[2] “Apa Dictionary of Psychology,” American Psychological Association. [Online]. Available: https://dictionary.apa.org/facial-expression. [Accessed: 09-Feb-2022].

[3] “ArXiv.” [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1203/1203.6722.pdf. [Accessed: 09-Feb-2022].

[4] “Papers with code - fer2013 dataset,” FER2013 Dataset | Papers With Code. [Online]. Available: https://paperswithcode.com/dataset/fer2013. [Accessed: 09-Feb-2022].

[5] “ArXiv:1307.0414v1 [stat.ML] 1 jul 2013.” [Online]. Available: https://arxiv.org/pdf/1307.0414v1.pdf. [Accessed: 09-Feb-2022].

[6] Learning social relation traits from face images. [Online]. Available: qqqq [Accessed: 09-Feb-2022].

[7] Follow meAjitesh KumarI have been recently working in the area of Data Science and Machine Learning / Deep Learning. In addition, “Machine learning - training, Validation & Test Data Set,” Data Analytics, 13-Jun-2021. [Online]. Available: https://vitalflux.com/machine-learning-training-validation-test-data-set/#:~:text=Generally%2C%20the%20training%20and%20validation,set%20aside%20for%20validation%20purposes. [Accessed: 09-Feb-2022].

[8] K. P. Shung, “Accuracy, precision, recall or F1?,” Medium, 10-Apr-2020. [Online]. Available: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9. [Accessed: 12-Apr-2022].

[9] J. Bechtel and J. Bechtel, “Two major concerns about the ethics of facial recognition in public safety,” Design World, 29-Oct-2019. [Online]. Available: https://www.designworldonline.com/two-major-concerns-about-the-ethics-of-facial-recognition-in-public-safety/. [Accessed: 09-Feb-2022].

[10] N. Martinez-Martin, “What are important ethical implications of using facial recognition technology in health care?,” Journal of Ethics | American Medical Association, 01-Feb-2019. [Online]. Available: https://journalofethics.ama-assn.org/article/what-are-important-ethical-implications-using-facial-recognition-technology-health-care/2019-02. [Accessed: 09-Feb-2022].