PURBANCHAL UNIVERSITY

DEPARTMENT OF COMPUTER ENGINEERING

KHWOPA ENGINEERING COLLEGE
LIBALI-2, BHAKTAPUR

A FINAL REPORT

ON

Real Time Facial Expression Recognition System

A Project work submitted for the partial fulfillment of requirements for the degree of Bachelor of Engineering in Computer Engineering (Eight Semester)

SUBMITTED BY

Ajay Shrestha (710301)

Anish Karmi (710303)

Sajan Basnet (710336)

Sudan Krishna Shrestha (710345)

Sushant Ghimire (710347)

UNDER THE GUIDANCE OF

Er. Reena Manandhar

19th November 2018

DEPARTMENT OF COMPUTER ENGINEERING

KHWOPA ENGINEERING COLLEGE
LIBALI-2, BHAKTAPUR

CERTIFICATE

This is to certify that the project entitled "**Real Time Facial Expression Recognition System**" submitted by Mr. Ajay Shrestha, Mr. Anish Karmi, Mr. Sajan Basnet, Mr. Sudan Krishna Shrestha & Mr. Sushant Ghimire in a partial fulfillment of the requirements for the award of the Degree of Bachelor of Engineering in Computer Engineering of Purbanchal University, is a bonafide work to the best of my/our knowledge and may be placed before the examination Board for their consideration.

Panel of Examiners:

Name	**Signature**	**Date**

External Examiner

Er……

Project Supervisor

Er. Reena Manandhar

Head of Department

Er. Reena Manandhar

ACKNOWLEDGEMENT

This project named **Real Time Facial Expression Recognition System** has been prepared under department of Computer Engineering. We are very grateful to our department for giving us an enthusiastic support.

It gives us pleasure to write down here the name of Er. Reena Manandar, Head of Department of Computer Engineering, Khwopa Engineering College as our supervisor for this project. We would like to thank her a lot for guiding properly throughout our project and continuously motivating to tackle the problems faced during this period. She has been always very much supportive and always motivating us to do best.

Any suggestion regarding the betterment of this project will be always welcomed.

Ajay Shrestha
Anish Karmi
Sajan Basnet
Sudan Krishna Shrestha
Sushant Ghimire

ABSTRACT

Facial expressions play an important role in interpersonal relations as well as for security purposes. The malicious intensions of a thief can be recognized with the help of his gestures, facial expressions being its major part. This is because humans demonstrate and convey a lot of evident information visually rather than verbally. Although humans recognize facial expressions virtually without effort or delay, reliable expression recognition by machine remains a challenge. A picture portrays much more than its equivalent textual description.

We will be developing a convolutional neural network model for classifying human emotion from dynamic facial expressions in real time. A large number of dataset will be used for training of our model to make it optimally accurate as possible.

Keywords: Neural Network, CNN, Facial Gestures, Face detection.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Abbreviation

CNN	Convolutional Neural Network
EEG	Electroencephalography
FERC	Facial Expression Recognition Challenge
FERS	Face Expression Recognition System
JAFFE	Japanese Female Facial Expression
LVQ	Learning Vector Quantization
RaFD	Radboud Faces Database
ReLU	Rectified Linear Unit
VGG	Visual Geometry Group

CHAPTER 1

INTRODUCTION

1.1 Background

One of the current top applications of artificial intelligence using neural networks is the recognition of faces in photos and videos. Most techniques process visual data and search for general patterns present in human faces. Face recognition can be used for surveillance purposes by law enforcers as well as in crowd management. Other present-day applications involve automatic blurring of faces on Google Street view footage and automatic recognition of Facebook friends in photos.

An even more advanced development in this field is emotion recognition. In addition to only identifying faces, the computer uses the arrangement and shape of eyebrows and lips to determine the facial expression and hence the emotion of a person. One possible application for this lies in the area of surveillance and behavioral analysis by law enforcement. Furthermore, such techniques are used in digital cameras to automatically take pictures when the user smiles. However, the most promising applications involve the humanization of artificial intelligent systems. If computers are able to keep track of the mental state of the user, robots can react upon this and behave appropriately. Emotion recognition therefore plays a key-role in improving human machine interaction.

There are seven types of human emotions shown to be universally recognizable across different cultures: anger, disgust, fear, happiness, sadness, surprise and contempt. On a day-to-day basis, humans commonly recognize emotions by characteristic features displayed as part of a facial expression. For instance, happiness is undeniably associated with a smile, or an upward movement of the corners of the lips. This could be accompanied by upward movement of the cheeks and wrinkles directed outward from the outer corners of the eyes. Similarly, other emotions are characterized by other deformations typical to the particular expression. Interestingly, even for complex expressions where a mixture of emotions could be used as descriptors, cross-cultural agreement is still observed. Therefore, a utility that detects emotion from facial expressions would be widely applicable. Such an advancement could bring applications in medicine, marketing and entertainment.

In this project we have been trying to build an artificial intelligent system that is capable of recognizing five basic emotions (happy, sad, angry, surprise, and neutral) of a person from their

facial expressions and for this we implemented convolutional neural network (CNN). We have used the FER-2013 dataset of labelled headshots for training our CNN model. We then transferred the skills learned by our model on static images into a real-time emotion recognition system, which continuously detects faces from a video feed and classifies the predominant emotion of the individual or a group of individuals.

1.2 Motivation

Human facial expressions can be easily classified into 7 basic emotions: happy, sad, surprise, fear, anger, disgust, and neutral. Our facial emotions are expressed through activation of specific sets of facial muscles. These sometimes subtle, yet complex, signals in an expression often contain an abundant amount of information about our state of mind. Through facial emotion recognition, we are able to measure the effects that content and services have on the audience/users through an easy and low-cost procedure. For example, retailers may use these metrics to evaluate customer interest. Healthcare providers can provide better service by using information about patient's emotional state during treatment. Entertainment producers can monitor audience engagement in events to consistently create desired content. So, these are the main source of motivation that made us want to build an application that gives machines the ability to make inferences about our emotional state.

1.3 Statement of Problem

It has often been said that the eyes are the "window to the soul." This statement may be carried to a logical assumption that not only the eyes but the entire face may reflect the "hidden" emotions of the individual. Darwin's research on facial expressions has had a major impact on the field in many areas; foremost, his belief that the primary emotions conveyed by the face are universal. Darwin placed considerable emphasis on the analysis of the action of different muscle groups in assessing expression. The research on the statement of Darwin was done by Ekman and Friesen. They hypothesized that the universals of facial expression are to be found in the relationship between distinctive patterns of the facial muscles and particular emotions (happiness, sadness, anger, fear, surprise, disgust and interest). They suggested that cultural differences would be seen in some of the stimuli, which through learning become established as elicitors of particular emotions, in the rules for controlling facial behavior in particular social settings and in many of the consequences of emotional arousal.

Many factors impinge upon the ability of an individual to identify emotional expression. Social factors, such as deception, and display rules, affect one's perception of another's emotional state. Therefore, there is a need to develop Face Expression Recognition System

1.4 Objective

- To build a real time system that can detect the emotions of a person by recognizing their facial expression.

1.5 Scope and Application

As human facial expression recognition is a very elementary process, it is useful to evaluate the mood or emotional state of a subject under observation. As such, tremendous potential lies untapped in this domain. The basic idea of a machine being able to comprehend the human emotive state can be put to use in innumerable scenarios, a few of which we have mentioned as follows:

- The ability to detect and track a user's state of mind has the potential to allow a computing system to offer relevant information when a user needs help – not just when the user requests help, for instance, the change in the Room Ambience by judging the mood of the person entering it.
- Clever Marketing is feasible using emotional knowledge of a patron and can be done to suit what a patron might be in need of based on his/her state of mind at any instant.
- Applications in surveillance and security. For instance, computer models can make correct classification of innocent or guilty participants based on the macro features extracted from the video camera.
- In this regard, lie detection amongst criminal suspects during interrogation is also a useful aspect in which this system can form a base. It is proven that facial cues more often than not can give away a lie to the trained eye.
- Patient Monitoring in hospitals to judge the effectiveness of prescribed drugs is one application to the Health Sector. In addition to this, diagnosis of diseases that alter facial features and psychoanalysis of patient mental state are further possibilities.

1.6 Limitation

This system has its own limitations. The limitations of this project due to various circumstances are as follows:

- The current face detection algorithm only detects faces which are nearly vertical.
- Current model was trained using only grayscale images which affects the classification performance and accuracy for real time video.

CHAPTER 2

LITERATURE REVIEW

Companies are already using sentiment analysis to gauge consumer mood towards their product or brand. By mining tweets, reviews, and other sources, companies can easily derive sentiment from natural language. Customer representatives in stores can see when a customer is frustrated or angry, but they can't be everywhere at once. However, companies have been using in-store cameras to monitor customer behavior for years. Companies use this data to track hot-spots in the store, the paths customers take, and even what they're specifically looking at. So this can also be used to monitor the customer's emotion. The primary issue is that it's difficult to translate facial muscles into emotions. It's easy for humans because we've had years of practice, but computers see the world as a grid of numbers that represent pixel values. We're able to look at an image of a person's face and easily differentiate between a smile and a frown, but for a machine learning model, it's a much more difficult task.

We're going to use a deep convolutional neural network to train the model. A convolutional neural network extracts features from 2D data and assigns weights to those features, eventually resulting in a prediction. In the case of facial emotion detection, the upward curves of a smile would be associated with happiness.

A Convolutional Neural Network (CNN) is comprised of one or more convolutional layers (often with a subsampling step) and then followed by one or more fully connected layers as in a standard multilayer neural network. The architecture of a CNN is designed to take advantage of the 2D structure of an input image. This is achieved with local connections and tied weights followed by some form of pooling which results in translation invariant features. Another benefit of CNNs is that they are easier to train and have many fewer parameters than fully connected networks with the same number of hidden units.

2.1 Related Work

A breakthrough publication on automatic image classification in general is given by Krizhevsky and Hinton. This work shows a deep neural network that resembles the functionality of the human visual cortex. Using a self-developed labelled collection of 60000 images over 10 classes, called the CIFAR-10 dataset, a model to categorize objects from pictures is obtained. Another important outcome of the research is the visualization of the filters in the network, such that it can be assessed how the model breaks down the pictures.

In another work which adopts the CIFAR-10 dataset [2], a very wide and deep network architecture is developed, combined with GPU support to decrease training time. On popular datasets, such as the MNIST handwritten digits, Chinese characters, and the CIFAR-10 images, near-human performance is achieved. The extremely low error rates beat prior state-of-the-art results significantly. However, it has to be mentioned that the network used for the CIFAR-10 dataset consists of 4 convolutional layers with 300 maps each, 3 max pooling layers, and 3 fully connected output layers. As a result, although a GPU was used, the training time was several days.

In 2010, the introduction of the yearly Imagenet challenge [3] boosted the research on image classification and the belonging gigantic set of labelled data is often used in publications ever since. In a later work of Krizhevsky et al., a network with 5 convolutional, 3 max pooling, and 3 fully connected layers is trained with 1.2 million high resolution images from the ImageNet LSVRC-2010 contest. After implementing techniques to reduce overfitting, the results are promising compared to previous state-of-the-art models. Furthermore, experiments are done with lowering the network size, stating that the number of layers can be significantly reduced while the performance drops only a little.

With respect to facial expression recognition in particular, Lv et al. [5] present a deep belief network specifically for use with the Japanese Female Facial Expression (JAFFE) and extended Cohn-Kanade (CK+) databases. The most notable feature of the network is the hierarchical face parsing concept, i.e. the image is passed through the network several times to first detect the face, thereafter the eyes, nose, and mouth, and finally the belonging emotion. The results are comparable with the accuracy obtained by other methods on the same database, such as Support Vector Machine (SVM) and Learning Vector Quantization (LVQ).

Another work on the Cohn-Kanade database [1] makes use of Gabor filtering for image processing and Support Vector Machine (SVM) for classification. A Gabor filter is particularly suitable for pattern recognition in images and is claimed to mimic the function of the human visual system. The emotion recognition accuracies are high, varying from 88% on anger to 100% on surprised. A big disadvantage of the approach however is that very precise preprocessing of the data is required, such that every image complies with a strict format before feeding it into the classifier.

One of the most recent studies on emotion recognition describes a neural network able to recognize race, age, gender, and emotion from pictures of faces [4]. The dataset used for the

latter category is originating from the Facial Expression Recognition Challenge (FERC-2013). A clearly organized deep network consisting of 3 convolutional layers, 1 fully connected layer, and some small layers in between obtained an average accuracy of 67% on emotion classification, which is equal to previous state-of-the-art publications on the same dataset. Furthermore, this thesis lays down a valuable analysis of the effect of adjusting the network size, pooling, and dropout.

CHAPTER 3

METHODOLOGY

3.1 Block Diagram

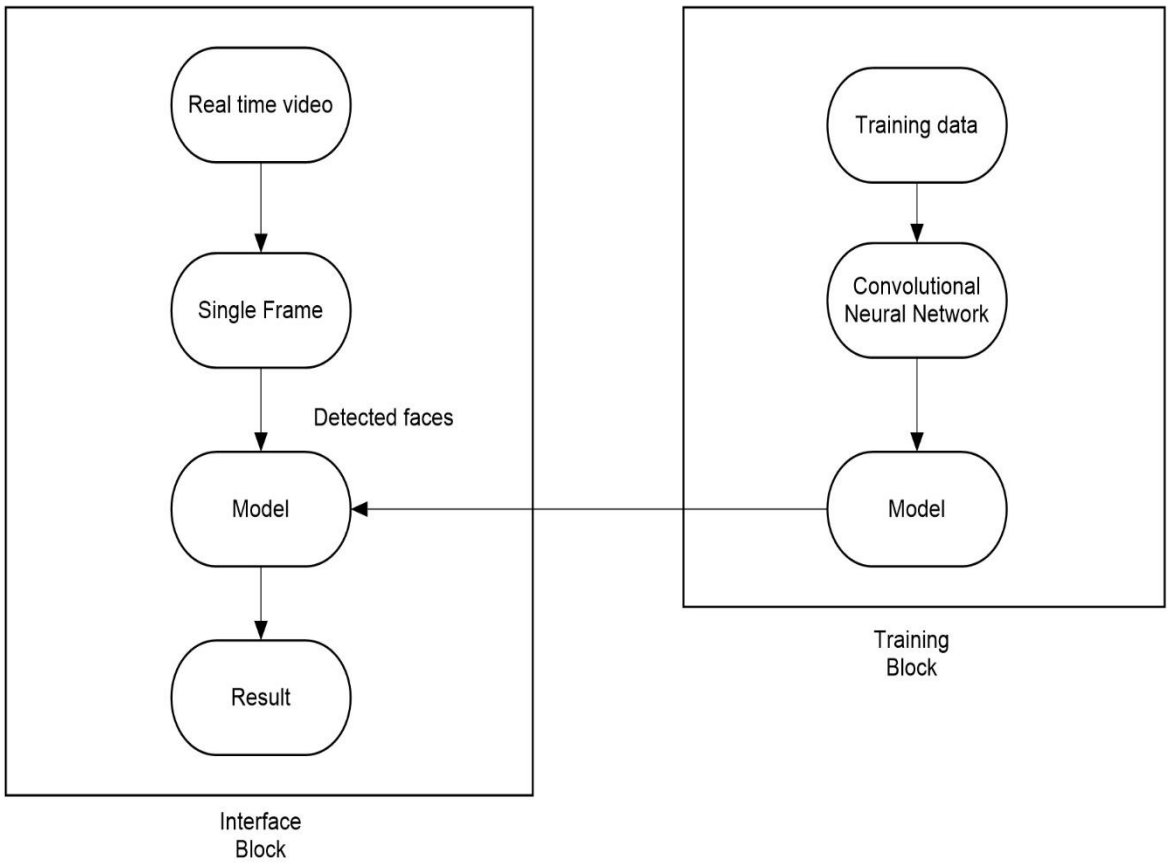

Figure 3.1: Block diagram

As shown in above Fig 3.2.1, a block diagram which represents our application. Our CNN model will be trained from the given dataset of static images. We then transferred the skills learned on static images into a real-time facial expression recognition system, which continuously detects, extract, crop, and grayscale the face region from a webcam video feed and classify the emotion of the person. During face extraction, the input image will be scaled and mapped to certain fixed size of (64*64) to fit the CNN.

3.2 Dataset

Neural networks, and deep networks in particular, are known for their need for large amounts of training data. Moreover, the choice of images used for training are responsible for a big part of the performance of the eventual model. This implies the need for a both high qualitative and quantitative dataset. For emotion recognition, several datasets are available for research, varying from a few hundred high resolution photos to tens of thousands smaller images. Some of the datasets available are the Facial Expression Recognition Challenge (FERC-2013), Extended CohnKanade (CK+), and Radboud Faces Database (RaFD).

The dataset we used for training the model is from a Facial Expression Recognition Challenge (FERC-2013). We used a total of 22323 pre-cropped, 48-by-48-pixel grayscale images of faces each labeled with one of the 5 emotion classes: anger, happiness, sadness, surprise, and neutral. The images in FER-2013 consist of both posed and unposed headshots and was created by gathering the results of a Google image search of each emotion and synonyms of the emotions.

Figure 3.2: Samples of the FER dataset.

3.3 The Model

Deep learning is a popular technique used in computer vision. We choose convolutional neural network layers as building blocks to create our model architecture.

3.3.1 Convolutional Neural Network

A typical architecture of convolutional neural network will contain an input layer, some convolutional layer, some dense layer (fully-connected layer) and an output layer as shown in figure below.

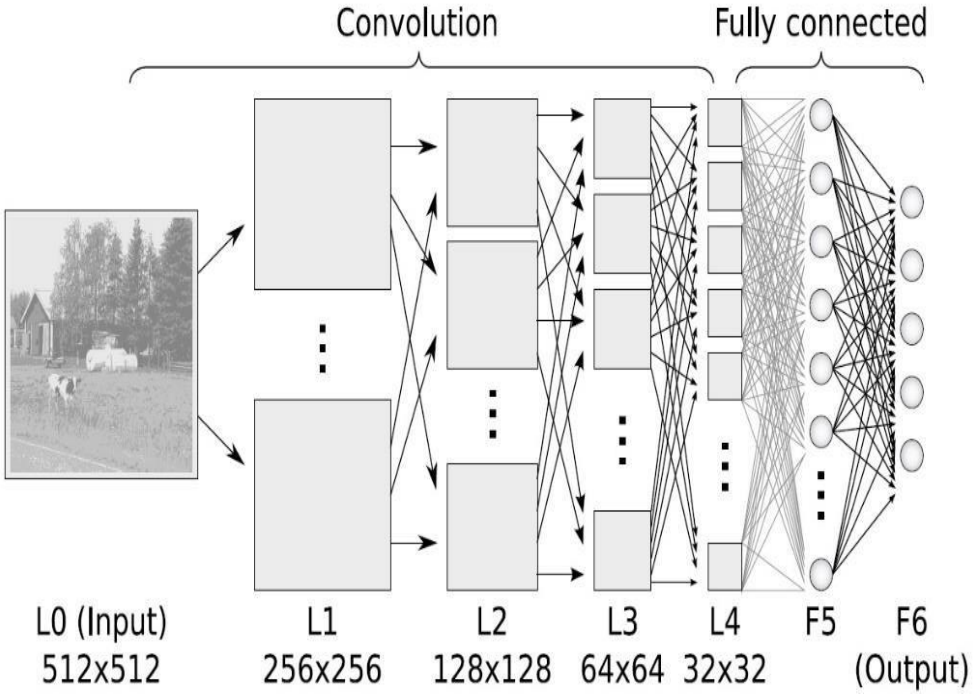

Figure 3.3: Layers in CNN

3.3.2 Input Layer

The input layer has pre-determined, fixed dimensions, so the image must be pre-processed before it can be fed into the layer. We used OpenCV, a computer vision library, for face detection in the image. The haarcascade_frontalface_default.xml in OpenCV contains pre-trained filters and uses Adaboost to quickly find and crop the face.

The cropped face is then converted into grayscale using cv2.cvtColor and resized to 64-by-64 pixels with cv2.resize. This step greatly reduces the dimensions compared to the original RGB format with three color dimensions (3, 64, 64). The pipeline ensures every image can be fed into the input layer as a (1, 64, 64) numpy array.

3.3.3 Convolutional Layer

The convolutional layer is the core building block of a CNN. The layer's parameters consist of a set of filters. During the forward pass, each filter is convolved across the width and height of the input, computing the dot product between the entries of the filter and the input and producing a 2-dimensional matrix of that filter. The input is converted into grayscale and passed through the convolutional layer. The filters used in this layer are Harr Cascade filter to detect edge and patterns.

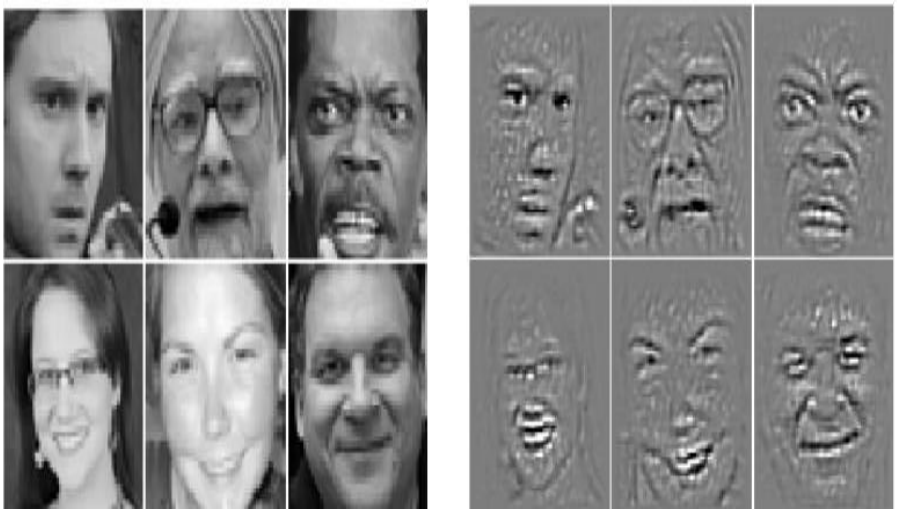

Figure 3.4: Output of filters

3.3.4 ReLU (Rectified Linear Unit)

A small but important layer in CNN is the Rectified Linear Unit or ReLU. Its math is also very simple-wherever a negative number occurs, swap it out for a 0. This helps the CNN stay mathematically healthy by keeping learned values from getting stuck near 0 or blowing up toward infinity.

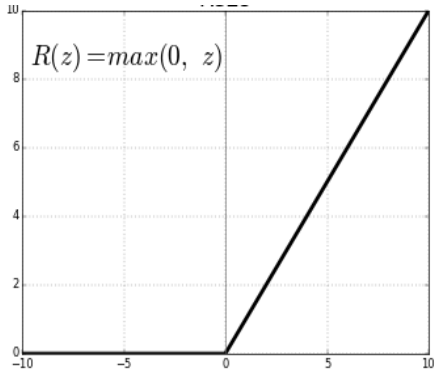

Figure 3.5: ReLU (Rectified Linear Unit)

The ReLU is half rectified (from bottom). f(z) is zero when z is less than zero and f(z) is equal to z when z is above or equal to zero. But the issue is that all the negative values become zero immediately which decreases the ability of the model to fit or train from the data properly. That means any negative input given to the ReLU activation function turns the value into zero immediately in the graph, which in turns affects the resulting graph by not mapping the negative values appropriately.

3.3.5 Pooling Layer

Pooling is a form of non-linear down-sampling. There are several non-linear functions to implement pooling among which max pooling is the most common. It partitions the input image into a set of non-overlapping rectangles and, for each such sub-region, outputs the maximum. The intuition is that the exact location of a feature is less important than its rough location relative to other features. We have used max pooling in our CNN which extracts sub regions of the feature map, keeps their maximum value, and discards all other values

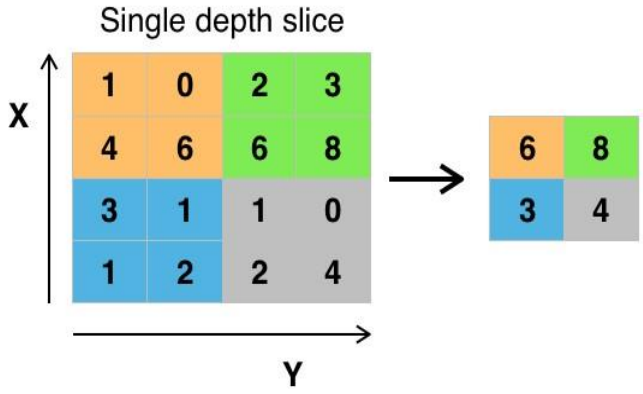

Figure 3.6: Pooling layer

3.3.6 Dense Layer

Finally, after several convolutional and max pooling layers, the high-level reasoning in the neural network is done via fully connected layers also known as dense layer. This layer perform classification on the features extracted by the convolutional layers and down sampled by the pooling layers. Neurons in a fully connected layer have connections to all activations in the previous layer, as seen in regular neural networks. Our final dense layer in a CNN contains a single node for each 5 emotions in the model, with a softmax activation function to generate a value between 0–1 for each node.

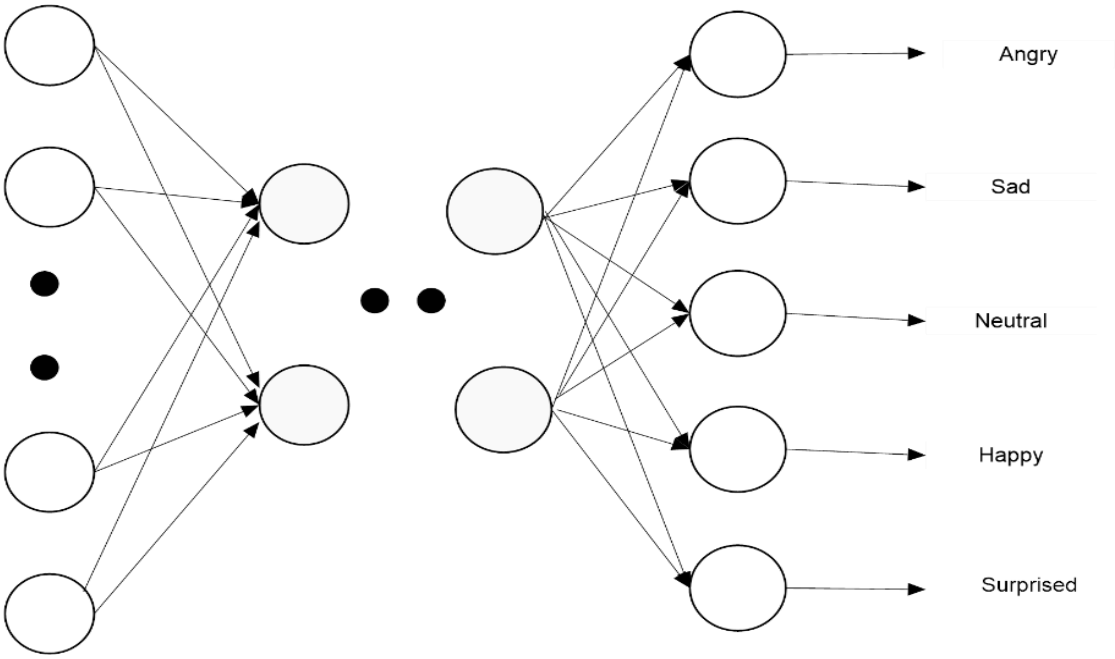

Figure 3.7: Dense layers

3.3.7 Output Layer

Instead of using sigmoid activation function, we used softmax at the output layer. This output presents itself as a probability for each emotion class. Therefore, the model is able to show the detail probability composition of the emotions in the face.

3.3.8 Softmax Activation Function

Sigmoid activation function is used for two class classification whereas softmax is used for multi class classification and is a generalization of the sigmoid function. In softmax, we get the probabilities of each of the class whose sum should be equal to 1. When the probability of one class increase then the probability of other classes decreases, so the class with highest probability is the output class.

For example, when predicting emotion in our application we may get output probabilities as 0.65 for happy, 0.20 for surprise , 0.7 for neutral, 0.5 for angry and 0.3 for sad in that case we take output with max probability as our final output. In this case it would be happy.

$$\sigma(x_j) = \frac{e^{x_j}}{\sum_i e^{x_i}}$$

Figure 3.8: Softmax Activation Function

3.4 Tools and Platforms

Following tools and platforms were used for developing this application

Software

- Text Editor: PyCharm, Python IDLE editor

Programming Language

- Python 3.6.5

Model Parameters

- Tensorflow 1.7.0
- Keras 2.1.5
- CNN model : mini_XCEPTION

Platforms

- Windows

CHAPTER 4

RESULT AND DISCUSSION

4.1 Neural Network Training

Convolutional Neural Net is a popular technique for current visual recognition tasks. We have created an application which can use the camera for real time data and be able to distinguish the emotion of the person in the frame. For the training, we used CNN and multiple hidden layers which will in term increase the accuracy of the application. The parameters of our model are given below:

Dataset used: FER2013

No. of data: 21577

Input size: 64*64

Filter window: 3*3

Training epochs: 500

Trained on: Colab GPU server

Training Time: 7 hours

Training Loss: 0.628959

4.2 Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or classifier) on a set of test data for which true value are known. Each cell is used to show the number of classified objects.

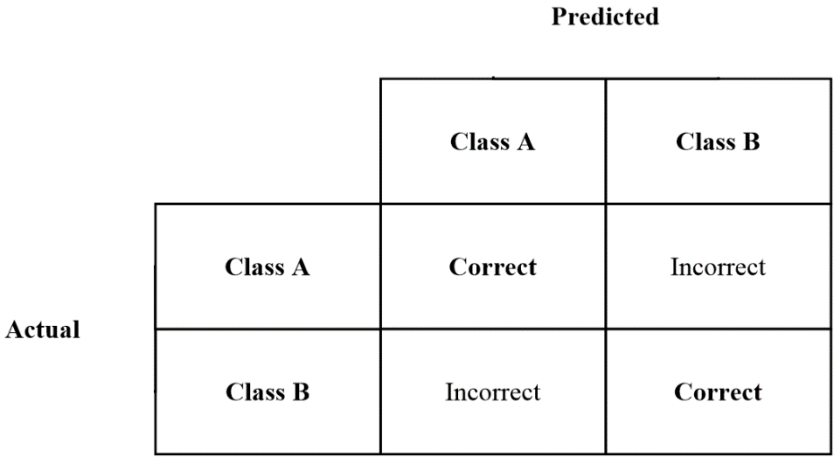

Figure 4.1: Confusion matrix example

The result of our model is shown in the following confusion matrix.

		Predicted				
		Angry	**Happy**	**Sad**	**Surprise**	**Neutral**
Actual	**Angry**	**2589**	127	352	63	329
	Happy	70	**6269**	73	96	206
	Sad	401	150	**2848**	33	694
	Surprise	60	116	26	**2503**	60
	Neutral	210	327	468	47	**3460**

Table 4.1: Confusion matrix of training

Accuracy: 81.888%

4.3 Test Cases

Test case 1:

		Predicted				
		Angry	**Happy**	**Sad**	**Surprise**	**Neutral**
Actual	**Angry**	**60**	5	6	3	14
	Happy	1	**104**	0	4	8
	Sad	9	3	**41**	1	18
	Surprise	1	1	1	**50**	1
	Neutral	3	2	5	0	**53**

Table 4.2: Confusion matrix of test case 1

Total images: 394 and Accuracy: 78.17%

Test case 2:

		Predicted				
		Angry	**Happy**	**Sad**	**Surprise**	**Neutral**
Actual	**Angry**	**39**	0	6	1	5
	Happy	2	**93**	0	3	2
	Sad	7	2	**61**	0	16
	Surprise	0	0	0	**44**	1
	Neutral	3	3	6	0	**58**

Table 4.3: Confusion matrix of test case 2

Total images: 352 and Accuracy: 83.80%

Test case 3:

		Predicted				
		Angry	**Happy**	**Sad**	**Surprise**	**Neutral**
Actual	**Angry**	**60**	6	8	4	17
	Happy	5	**203**	4	4	13
	Sad	13	9	**82**	6	31
	Surprise	2	5	1	**81**	4
	Neutral	9	20	27	4	**97**

Table 4.4: Confusion matrix of test case 3

Total images: 715 and Accuracy: 73.14%

4.4 Discussion

On the FER-2013 dataset, we achieve a test accuracy of 81.88% with our CNN model. Due to the nature of real-time classification, it is hard to get a definitive metric of our real-time system's accuracy. Our system reliably classified some emotions (the most reliable classification being happy), but struggled when lighting conditions changed, or backgrounds were noisy. Though we tried training on FER-2013 to better reflect real-time data, conditions like lighting in real-time differed from static images, making it hard for our program to transfer over skills learned on static databases.

CHAPTER 5

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

We created an application using Convolutional Neural Network to determine the emotion of a person in real time. Our application uses the webcam to scan for faces in the frame and captures the face coordinates. This is then used to crop the face and resize it to be analysed by our trained model. The model then determines the emotion of the face and shows it to the user around the box of the original face. Similarly, for multiple faces in a single frame, each face will be shown their emotions respectively.

5.2 Future Recommendation

For continued work on this project, we believe there are two major areas of focus that would improve our real time facial expression recognition system. First would be fine tuning the architecture of the CNN used for the model to fit perfectly with the problem at hand. Some examples of this fine tuning include finding and removing redundant parameters, adding new parameters in more useful places in the CNN's structure, adjusting the learning rate decay schedule, adapting the location and probability of dropout and experimenting to find ideal stride sizes. Also, with the use of more complex face detection algorithm, facial expression can be detected more precisely.

A second area of focus lies in adapting the datasets to more closely reflect real-time recognition conditions. For example, simulating low light conditions and "noisy" image backgrounds, could help the model become more accurate in real-time recognition. Additionally, making sure that the distribution of models in the training dataset accurately reflects the distribution of subjects that the system will see when running in real-time.

REFERENCES

[1] T. Ahsan, T. Jabid, and U.-P. Chong. Facial expression recognition using local transitional pattern on gabor filtered facial images. IETE Technical Review, pages 47-52, 2013.

[2] D. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 3642-3649. IEEE, 2012.

[3] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 248-255. IEEE, 2009.

[4] A. Gudi. Recognizing semantic features in faces using deep learning. arXiv preprint arXiv:1512.00743, 2015.

[5] Y. Lv, Z. Feng, and C. Xu. Facial expression recognition via deep learning. In Smart Computing (SMARTCOMP), 2014 International Conference on, pages 303-308. IEEE, 2014.

[6] Google colaboratory
https://colab.research.google.com

[7] FER dataset (date: 2018-06-11 time: 18:30),
Available: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data

APPENDIX

```
[[ 60.    6.    8.    4.   17.]
 [  5. 203.    4.    4.   13.]
 [ 13.    9.   82.    6.   31.]
 [  2.    5.    1.   81.    4.]
 [  9.   20.   27.    4.   97.]]
Total no. of images 715.0
Total no. of correct emotion 523.0
Accuracy: 0.7314685314685314
```

Figure A.1: Confusion matrix

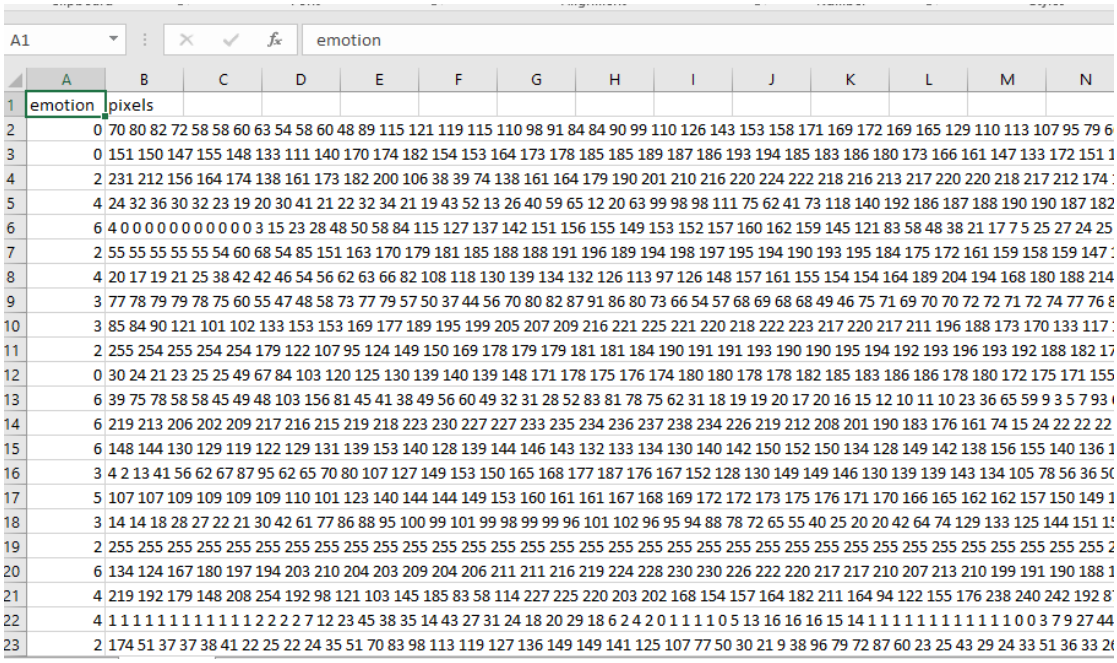

Figure A.2: Dataset

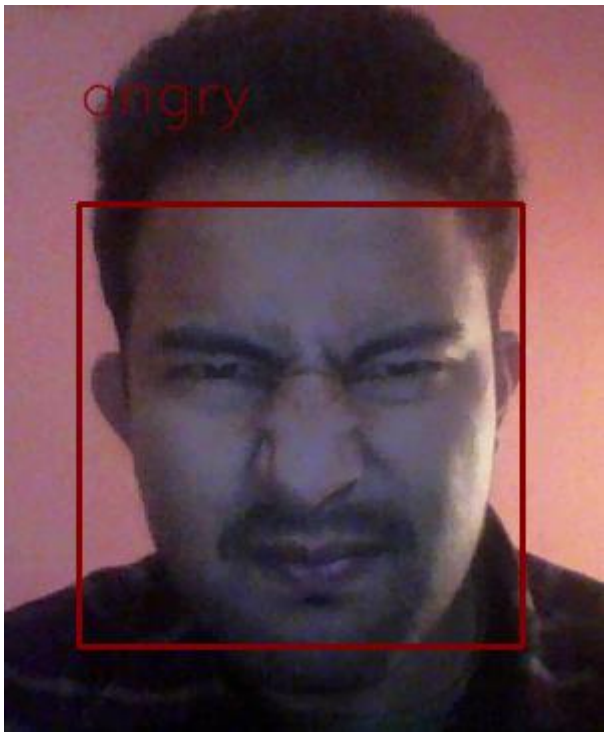

Figure A.3: Angry face

Figure A.4: Happy face

Figure A.5: Sad face

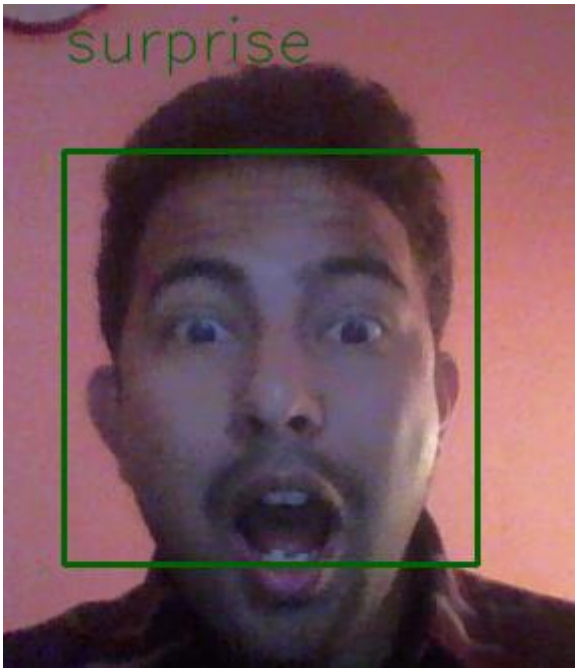

Figure A.6: Surprise face

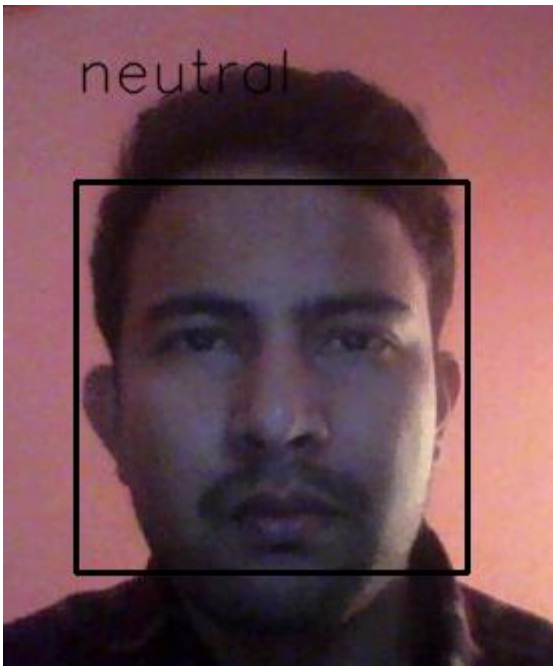

Figure A.7: Neutral face

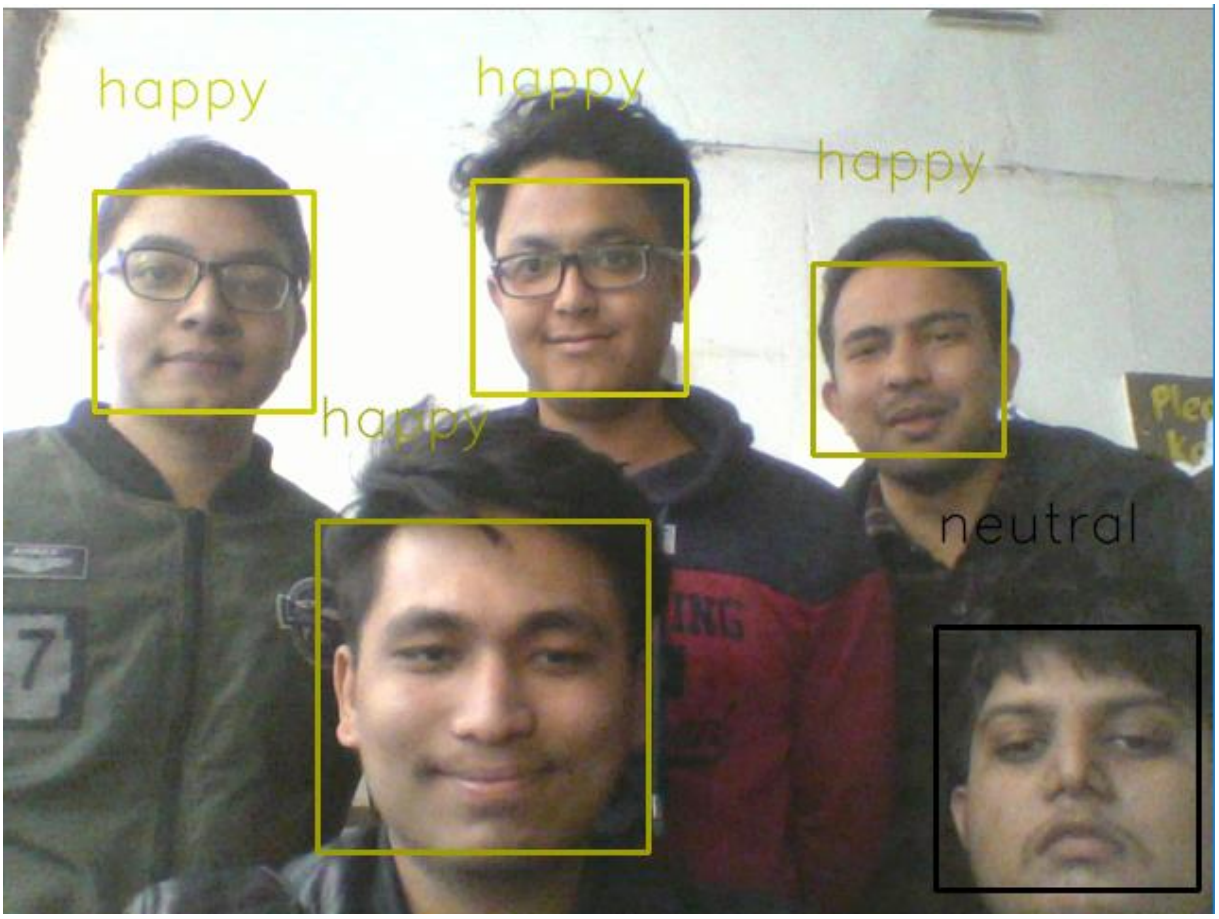

Figure A.8: Multiple faces