****

**Sentence-Level Emotion Apprehension Through Facial Expression & speech verification Analysis**

**Thesis**

**Submitted By**

|  |  |
| --- | --- |
| 17-33833-1 | Md. Mohaimanul Haque |
| 17-33872-1 | Souvik Das Dipta |
| 17-33658-1 | Abu Fuzail polin |
| 17-33602-1 | Ashik Al Habib |

**Department of Computer Science**

**Faculty of Science & IT**

**American International University Bangladesh**

**February, 2021**

|  |
| --- |
| **Declaration** |

We declare that this thesis is our original work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

|  |  |
| --- | --- |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Md. Mohaimanul Haque**  17-33833-1  Faculty of Science & Technology | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Souvik Das Dipta**  17-33872-1  Faculty of Science & Technology |
|  |  |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Fuzail polin**  17-33658-1  Faculty of Science & Technology | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Ashik Al Habib**  17-33602-1  Faculty of Science & Technology |

|  |
| --- |
| **Approval** |

The thesis titled “Sentence-Level Emotion Apprehension Through Facial Expression & speech verification Analysis” has been submitted to the following respected members of the board of examiners of the department of computer science in partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science on (18 February, 2021) and has been accepted as satisfactory.

|  |  |
| --- | --- |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **VICTOR STANY ROZARIO**  Assistant Professor & Supervisor  Department of Computer Science  American International University-Bangladesh | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **External’s Name**  Rank & External  Department of Computer Science  American International University-Bangladesh |
|  |  |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Dr. Dip Nandi**  Asst. Professor & Head (Undergraduate)  Department of Computer Science  American International University-Bangladesh | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Professor Dr. Tafazzal Hossain**  Dean  Faculty of Science & Information Technology  American International University-Bangladesh |
|  |  |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Dr. Carmen Z. Lamagna**  Vice Chancellor  American International University-Bangladesh | |

|  |
| --- |
| **Acknowledgement** |

First and foremost, praise and thanks to God, the Almighty for his showers of blessings throughout my research work to complete the research successfully.

We would like to thank the Department of Computer Science -American International University-Bangladesh for including THESIS course into our course curriculum. We have achieved so much experience and knowledge throughout completing the whole research.

We would like to express our sincere gratitude to our Thesis advisor VICTOR STANY ROZARIO, Assistant Professor -Department of Computer Science, American International University-Bangladesh for his patience, motivation and immense support. His guidance helped us throughout the research and writing of this thesis.

Last but not the least we would like to thank our parents for supporting and trusting us during our whole life. Their hard work for educating and preparing us for our future made us mentally strong and made us capable of doing something great in life.

|  |
| --- |
| **Abstract** |

The importance of Emotional state apprehension is widely perceived in social interaction and social intelligence. This analysis has been an active research topic since the 19th century [10]. In human-to-human communication, the understanding of facial expressions forms a communication carrier that offers vital data about the mental, emotional and even physical state of the persons in conversation [10]. Inevitably user's emotional state plays an important role not only in human associations with other people but also in the way a user uses computers. As the emotional state of a person may determine consistency, task solving, and decision-making skills [12]. Through this study, facial expression analysis refers to computer systems that attempt to automatically predict user emotional state by analysing and recognizing facial motions and facial feature changes from visual information. Though Interpretation is aided by circumstances, body gesture, voice, individual diversity, and cultural factors as well as by facial arrangement and timing [6]. In this study, facial expression analysis methods will be conducted to analyse the facial actions notwithstanding of context, culture, gender, and so on.

|  |
| --- |
| **Table of Contents** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chapter 1: Introduction** | | | **09** |
| 1.1 | Background of the study | | 09 |
| 1.2 | Statement of the problem. | | 10 |
| 1.3 | Goals and objective. | | 10 |
|  |  | |  |
| **Chapter 2: Literature Review** | | | **11** |
| 2.1 | Convolutional neural network | | 11 |
| 2.2 | Issues with conventional method | | 12 |
| 2.3  2.4  2.5 | Technical challenges.  2.3.1 Data  2.3.2 Face occlusion and lighting issues  2.3.3 Recognizing incomplete emotions  CNN  Speech process  2.5.1 Stop words  2.5.2 Tokenization | | 12  12  13  13  13  15  15  16 |
| **Chapter 3: Methodology** | | | **17** |
| 3.1 | Work Flow | | 17 |
| 3.2 | Video Processing | | 18 |
|  | 3.2.1 | Video to frames. | 18 |
|  | 3.2.2  3.2.3  3.2.4 | Video to audio  Reducing noise  Audio to text | 19  19  19 |
| 3.3 | Model Creation | | 19 |
| 3.4 | Training | | 21 |
| 3.5 | Emotion Recognition from Images | | 21 |
|  | 3.5.1 | Image processing | 21 |
|  | 3.5.2 | Prediction | 21 |
| 3.5 | Emotion Recognition form Audio | | 21 |
|  | 3.6.1  3.6.2 | Removing stop words  Prediction | 21  22 |
| **Chapter 4: Result and Discussion** | | | **23** |
| **Chapter 5: Future Works** | | | **26** |
| **Chapter 6: References** | | | **27** |

|  |
| --- |
| **List of Figures** |

|  |  |  |
| --- | --- | --- |
| **Fig. 2.4** | Convolutional Neural Network | 13 |
| **Fig. 2.4.1** | Local Receptive Field | 14 |
| **Fig. 2.4.2** | Max-Pooling | 14 |
| **Fig. 2.4.3** | SoftMax Function | 14 |
| **Fig. 3.1** | Work Flow (Emotion Recognition from Text) | 16 |
| **Fig. 4.0** | Frames taken from the video | 21 |
| **Fig. 4.1** | Results extracted from the video | 22 |
| **Fig. 4.2** | Video Recognition | 22 |
| **Fig. 4.3** | Pi-Chart | 23 |
| **Fig. 4.4** | Merged Final Result | 23 |
| **Fig. 4.5** | Work Flow (Emotion Recognition from Image) | 16 |
|  |  |  |

|  |
| --- |
| **Chapter 1: Introduction** |

**1.1 Background of the study**

This region includes a comprehensive survey on the approaches previously used to detect the emotion of a sentence. It can be seen that Dipankar Das and Shivaji Bandyopadhyay [4] operate a machine learning approach using Conditional Random Field (CRF). Their method is a two-step method where the first step is to create an emotion for each word in a sentence using WordNet’s Affect List and the next step is to find the authoritative sentiment of each sentence using weight points of each word in the sentence. The first step uses CRF for word-level commentary which searches for words in the sentence that are present in the Affect list and returns phrase tagging of emotions. The next step is performed by applying these word-level emotions to achieve the overall sentence level emotion based on weight scores of each word in a sentence. This paper produces a high accuracy of 87.65%, it is limited by the Synonyms set (SynSet) in the Affect list. This model failed to consider words that do not exist in the said list. The scholars in Paper [5] do an experimental analysis of different approaches to tackle the problem of sentence-level emotion tagging. They use five approaches, which include four knowledge-based ideas and one corpus-based idea. The first approach in knowledge-based ideas, WordNet-Affect Presence, deals with interpreting the disturbances in a text simply based on the residence of words from the Word-Net Affect lexicon. wherein the second approach, Latent Sentiment Analysis (LSA), the LSA similarity between the given text and each emotion is determined and each emotion is defined as a vector of the word expressing that particular emotion. In the third method, its synonyms from the WordNet SynSet are used. Ultimately, the fourth scheme, LSA all emotion words, serves the preceding set by combining the words in all the SynSets identified with a given sentiment, as found in WordNet Affect List. Besides, the corpus-based procedure used in this paper [5] uses a machine-learning classifier, Naive Bayes that is trained on Blog Posts to incorporate emotion in a labeled data set. This approach is more practical and has been employed similarly in our model. In paper [6], the authors discuss a design to find emotion labels using two methods: keyword spotting and lexical affinity. Certain methods use the existing lexical corpus to find words about a distinct emotion of the stated collection of emotions. In paper [6] negation was not considered. words like not, neither, and never, which can give the polarizing emotion of the sentence. it tags an emotion to sentence based on the context preferably than the word level emotion weights. This process assigns an emotion to the sentence by weighing the relation between the different words and emotions present in it. the ANEW list (Bradley and Lang 1999) is advised, and if found, its purpose is obtained.

**1.2 Statement of the problem**

The market for emotion-detection technology is worth approximately $21.6bn, and its value is prognosticated to more than double by 2024, relinquishing $56bn. Businesses can purchase systems to help them vet job applicants, analyze advertisements for their sentimental impression and test criminal defendants for signs of fraud [19]. The facial expression identification system was introduced in 1978 by Suwa et. al [10]. manifested an attempt to automatically interpret facial expressions by tracking the motion of 20 recognized points on an image series in 1978. The influence of the facial expression system is broadly acknowledged in social interaction and social intelligence. Facial expression analysis includes both estimations of facial motion and recognition of expression.

Paul Ekman and Friesen carrying out more in-depth Darwin studies have concluded that there are six transcultural and prototypic emotional expressions [7]. These basic expressions are happiness, anger, sadness, surprise, disgust, and fear. Some authors consider the neutral face as a seventh expression [7]. It is known that facial expression changes through a set or subset of prototypic emotional expressions. Given the huge influence, Ekman’s work possessed on the science of emotion, his opinions are pervasive in emotion-detection technology, informing many algorithms, including those sold by Microsoft, IBM and Amazon.

Ekman’s study is questionable not just because it underplays the importance of developmental differences, but also because it considers that there is a relationship between someone’s facial expression and their emotional state. Researchers have since discovered the exact opposite: a recent study by the Ohio State University implied that facial appearances are often unpredictable signs of emotion.

**1.3 Goals and Objective**

Our main goal focuses on achieving a model that can segregate the basic emotions: angry, sad, surprise, disgust, happiness, neutral, and fear and to obtain the accuracy better than the standard 14.29%. We are also determined to analyze the result of our model in terms of the accuracy of each represented class. Further development of the model may be expected to achieve further detailed classification that has a more complex variance of the condition than lab condition images. We are determined to further develop the system that might include the frequency of human speech as well as gesture recognition.

|  |
| --- |
| **Chapter 2: Literature Review** |

In the Emotion detection research field, several domains contribute like machine learning, natural language, neuroscience, etc. In previous researches, facial expression, voice features, or textual data are used separately to classify emotions. Emotion can be classified into several static classifications like happiness, sadness, disgust, anger, fear, and surprise [20]. It can be further improved by combining the image, voice, and textual data. This combination of data gives further improved results.

**2.1 Convolutional Neural Network:**

Like every other classification problem, the emotion recognition problem requires an algorithm to complete feature extraction and categorical classification. In order to classify an emotion, we need to extract certain features from data and build a model that can classify the input based on the feature. The procedure can be outlined as follows:

**1. Data Pre-processing:**

The data pre-processing is to standardize the data. The typical way is to set the mean to 0 and to also divide the data by the standard deviation.

**2. Feature Extraction:**

The typical method is to detect the face and extract the Units from the face, and certain emotions contain the combination of AUs code [21].

**3. Model Construction:**

The conventional classifier can be either a supervised or unsupervised algorithm. A typical example of a supervised algorithm is Support Vector Machine [5], and the examples of the unsupervised algorithm include Principal Component Analysis (PCA) and Linear Discriminant Analysis.

**4. Label or Result Generation:**

The typical way to generate a label or result is to find which decision boundary has the minimum Euclidean distance from the data.

**2.2 The issues with conventional method:**

Classification of facial expression can sometimes be difficult, surprisingly for humans also. It is shown on several researches that people might recognize different emotions for the same facial expression. And it’s even harder for AI. A lot of factors make emotion recognition tricky. We can divide these factors as follows: technical and psychological.

**2.3 Technical challenges:**

Emotion detection has many issues detecting moving object in the video: detecting a target, uninterrupted detection, deficient or unforeseeable actions and many more. We have to overcome these challenges to obtain optimum model for our work.

The emotion detection model examine facial features like eyes, noses, eyebrow position, chins, mouths and other features as actuation points. Occasionally, this detection intricate due to:

**The distance between features**: There might be some functionality in the system that treats the average distance between landmarks as an epitome and compare them only within the range.

**Feature size**: Sometimes the model might encounter some issues with detecting irregular features, like slender or pale lips, narrow eyes and many more.

**Skin color:** In some cases, a solution may misclassify a feature due to skin color.

Some paper indicates that implementing a part based model that divides facial features into many different sections according to the physical structures of the face. Then the model feeds the features into the networks with applicable labels [22].

**2.3.1 Data:**

When using machine learning or deep learning algorithms problems requires large amount of training data. Our data has videos with inconsistent frame rate, various angles and backgrounds, people of different genders and nationalities.

**Training Dataset:**

The training data (fer2013) contains 48x48 pixel grayscale images of different faces. The data has been automatically registered so that the faces are nearly centered and occupies similar space in each image.

**Test Dataset:**

For the Test Dataset we have used CMU-MOSEI multimodal Dataset. We had to pre-process the video file according to our needs.

**2.3.2 Face occlusion and lighting issues:**

Occlusion happens when the target’s posture is changed. This is frequent problem in motion detection in a video, especially if the video is raw or unprepared. We can dissipate the problem by formalizing the system that is responsible for detecting the facial features in the video, identifying the features and deduce them to a 3D model of human face.

We can further increase the efficiency of the model by introducing the illumination normalization algorithms.

**2.3.3 Recognizing incomplete emotions:**

Most algorithms focus on recognizing the high-intensity expression. This might results in an inconsistent recognition of emotions when examining people from cultures with traditions of emotional suppression.

**2.4 CNN:**

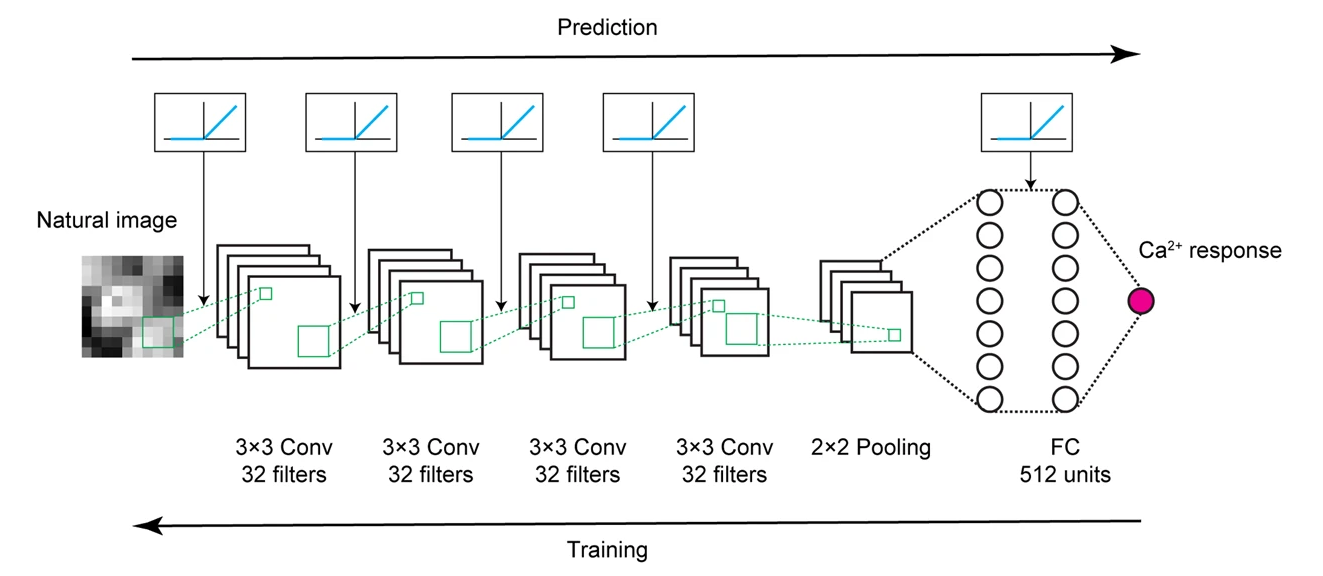
CNN model mainly contains 3 different substructures, which includes: convolutional layers, pooling layers, fully connected layers [23].

Figure 2.4: Convolutional Neural Network

Convolutional layer consists of many features maps. Local receptive field might be slithered when creating a new feature map over the input [24]. Different variations of data are used including images, text when performing convolution. Individual neuron in the layers might not be connected to every nodes or neurons in the preceding layer rather they are just associated to nodes in a special area known as the local receptive field.

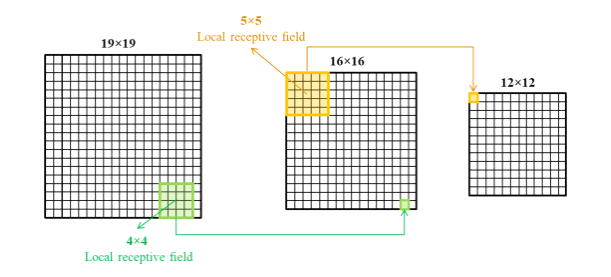


Figure 2.4.1: Local receptive field

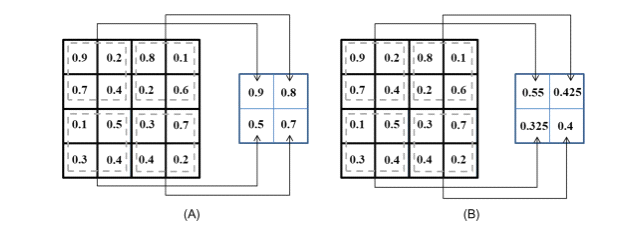
Pooling layers are initiated to simplify the information. These layers are also known as the subsampling layer. Pooling can be implemented in different types of problems such as harmonic average, geometric average, maximum pooling. Max-pooling and average-pooling are the most frequent processes for pooling. These layers are mandatory to decrease the computational time.

Figure 2.4.2: Max-pooling

Fully connected layers are the concluding layers in the CNN model, placed after the series of convolution layer and pooling layers. This layer allows the feature vector the input data that further used for some machine learning problems including prediction, classification [23]. The last layer of fully connected layers is also known as SoftMax Classifier.



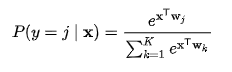


Figure 2.4.3: SoftMax function

The max-pooling layer performs the task to down sample the image or feature map, thereby reducing the computation and directing the next layers to focus on more detailed features. In conclusion, these 3 types of layers construct the uniqueness of the convolution block.

In addition to the convolutional block, the fully connected layers will serve as a classifier. Each unit of the layer contains the weight matrix, and through the linear transformation and activation function, the output becomes the input of the next layers of units [9]. In contrast to a traditional linear transformation, the activation function ReLU (shown in equation 4) will act in the same way as the convolutional layers to make the system more easily distinguish the feature. Output = max (input, 0)

In addition, the last layer is normally a SoftMax classifier, and the result is typically the one with maximum probability in multinomial distribution.

classI ∼ Multinomial(φi)

label = arg max.

**2.5: Speech Process:**

There are two types of data for computer- structured and unstructured. The structured data is highly organized and structured for processing and analysis within databases or spread sheets. Computers handle structured data very well. However, dealing with unstructured data or data that does not have a pre-defined structure or format, such as human language, is a challenge for them.

**2.5.1: Stop words:**

Words that are super common and doesn't carry that much of a meaning, they just connect the important words of a sentence are called stop-words. For example, in English "a", "the", "are", "is" etc. are very common in pretty much every English sentence. So, the prescience is that by removing these words, one can focus on the words that carry more prominence in a sentence or carry more information about the overall corpus. Different languages have different stop words. The Bengali language is a grammatically complex language. It has stop-words that needs to be removed for extracting meaning from a corpus. This technique commonly used in topic extraction, keyword searching, NLP classification tasks and so on.

**2.5.2: Tokenization:**

Tokenization is the method of splitting a text stream into symbols of words phrases and other significant items called tokens of symbols may be individual phrases of words or even whole phrases. Tokens can be phrases, individual words or even entire sentences. Any characters that are not words like punctuation marks, are discarded in the process of tokenization. This tokenized word becomes input for things such as parsing and text mining.

Thus we have used Text2Emotion for recognition of the context. Text2emotion is the python package developed with the clear intention to find the appropriate emotions embedded in the text data. The research concludes that when a person is in the thinking process and is clear approximately his statement then he will express his emotions in the right context of manner. Therefore it will be properly aligned. It gives an output as a dictionary labeling context into 5 basic emotion categories such as **Happy**, **Angry**, **Sad**, **Surprise,** and **Fear**.

|  |
| --- |
| **Chapter 3: Methodology** |

**3.1 Work Flow:**

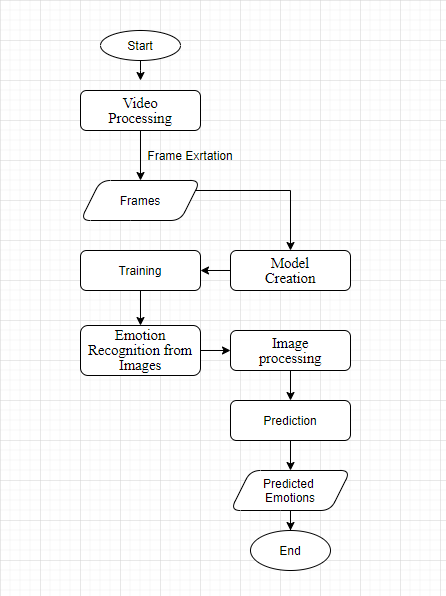
****

Figure 4.5: Work Flow (Frame recognition)

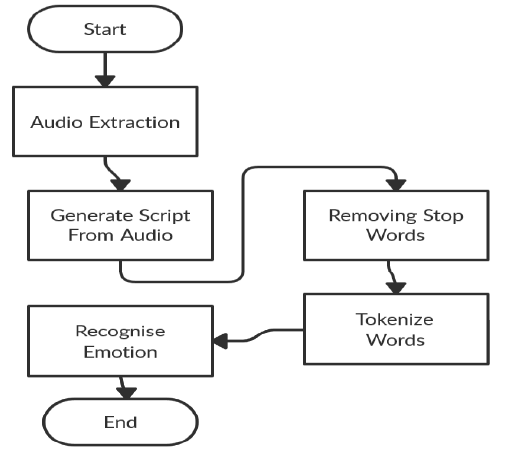


Figure 3.1: Work Flow (speech Recognition)

**3.2 Video Processing:**

The first step is to process the video and break the video into two part

**3.2.1 Video to Frame:**

First, we extract frame from the mp4 video which is located in Google drive. For extracting frame from video, we imported scikit-image, OpenCV-python library. First challenge is to have consistent data. Main rules in the dataset creation were:

* light conditions: records during the day
* video quality at least 640 width or above
* removed any cover with titles
* Face must be in the middle

Then we make each frame grayscale and resize it to 640. After that we stored every frame in Google Drive.

Algorithm:

count = 0

success = True

vidcap = cv2.VideoCapture(videofile)

while success:

    if (count%one\_frame\_each == 0):# checks frame number and keeps one\_frame\_each

        success,image = vidcap.read() # reads next frame

        image\_gray = rgb2gray(image)  # grayscale image

        if image.shape[1]>640:     # if image width > 640, resize it

          tmp = resize(image\_gray, (math.floor(640 / image\_gray.shap e[1] \* image\_gray.shape[0]), 640),mode='constant')

        plt.imsave("%s/%s%d.png" % (OUTPUT\_FRAMES\_PATH,frame\_name, count), tmp, cmap= plt.cm.gray) # saves images to frame folder

        print ('\*', end="")

        if count>=100:   #limits frames to 100

          break

    else:

        success,image = vidcap.read()# reads next frame

    count += 1

**3.2.2 Video to Audio:**

Initially, we Have extracted audio from Video using MoviePy. MoviePy can edit all the several popular audio and video forms, including GIF, and runs on Windows/Mac/Linux, with Python 2.7+ and 3.

**3.2.3 Reducing Noise:**

The default feature given by speech\_recognition which is adjust\_for\_ambient\_noise was used to reduce the noise of the audio.

**3.2.4 Audio to Text:**

we have used and compared multiple voice recognition technology to achieve the best outcome. The technologies include google Text to speech, Microsoft Bing Speech, Wit, SoundHound for generating the script. On a given audio file, Google, Wit had shown similar results which are about 45% of the whole audio was extracted correctly where Bing showed comparatively better output about 61% of the whole transcript correctly. But finally, Soundhound was our API of choice for recognizing speech which is 80%+.

**3.3 Model Creation:**

For creating the model we use the Sequential API from keras library. Our Model contains four Convolutional Layers, two dense layer and one hidden layer.

Convolutional Layers:  
First Convolutional layer has 64 filters, size is 3x3, output matrix and input matrix are same. Second layer contains 128 filter, third layer contains 256 filters and last layer contains 512 filters. Other parameters were same.

Model is given bellow:

def createModel():

model = Sequential()  
model.add(Conv2D(64, (3, 3), padding='same', input\_shape=(48,48,1)))

model.add(BatchNormalization())

model.add(Activation('relu'))

model.add(MaxPooling2D(pool\_size=(2, 2), strides=None, padding='same'))

model.add(Dropout(0.25))

model.add(Conv2D(128, (3, 3), padding='same'))

model.add(BatchNormalization())

model.add(Activation('relu'))

model.add(MaxPooling2D(pool\_size=(2, 2), strides=None, padding='same'))

model.add(Dropout(0.25))

model.add(Conv2D(256, (3, 3), padding='same'))

model.add(BatchNormalization())

model.add(Activation('relu'))

model.add(MaxPooling2D(pool\_size=(2, 2), strides=None, padding='same'))

model.add(Dropout(0.25))

model.add(Conv2D(512, (3, 3), padding='same'))

model.add(BatchNormalization())

model.add(Activation('relu'))

model.add(MaxPooling2D(pool\_size=(2, 2), strides=None, padding='same'))

model.add(Dropout(0.25))

model.add(Flatten())

model.add(Dense(512))

model.add(BatchNormalization())

model.add(Activation('relu'))

model.add(Dropout(0.25))

model.add(Dense(256))

model.add(BatchNormalization())

model.add(Activation('relu'))

model.add(Dropout(0.25))

model.add(Dense(7))

model.add(Activation('softmax'))

return model

model = createModel()

**3.4 Training:**

In case of Recognition from the Video frames, the task is to categorize each face based on the emotion shown in the facial expression into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The training set consists of 28,709 approx. examples and the public test set consists of 3,589 approx. examples.

In case of Audio Recognition, the task is to categorize each word based on the meaning into one of five categories on the scale 0 to 1. Which are: Angry, fear, Happy, Sad and Surprise.

We train the model for 30 epoch and batch size of 8. Which game the model to have 71.5% accuracy.

**3.5 Emotion Recognition from Images:**

For recognizing the images that were saved during video to frame conversion need to have some modification to fit into the model.

**3.5.1 Image Processing:**

Firstly we need to detect the faces from the images. For face detection we use python keras.preprocessing library. Then the images were converter to have 48x48 shape.

**3.5.2 Prediction:**

We used model.predict() function to predict the Emotion of the processed images taken form the videos. The predict() function takes an array of one or more data instances and enables us to predict the labels of the data values on the basis of the trained model. Then we take the max index of the predicted data and select the emotions.

**3.6 Emotion Recognition from Audio:**

**3.6.1 Removing Stop words:**

Words that are super common and doesn't carry that much of a meaning, they just connect the important words of a sentence are called stop-words. These should be removed too optimize and reduce valuable processing time. Here We have user NLTK (Naturak Language Toolkit) to Remove stop words from the Extracted Context.

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]

Thus, the prescience is that by removing these words, one can focus on the words that carry more prominence in a sentence or carry more information about the overall corpus.

**3.6.2 Prediction:**

We have used Text2Emotion python package to recognize the Emotion from the processed text extracted from the given video. Text2Emotion is a python package which automatically process texts, tokenize and extracts the emotions (Angry, fear, Happy, Sad and Surprise)

|  |
| --- |
| **Chapter 4: Result & Discussion** |

The Video we used, was compiled into frames of every second and text was extracted from video.

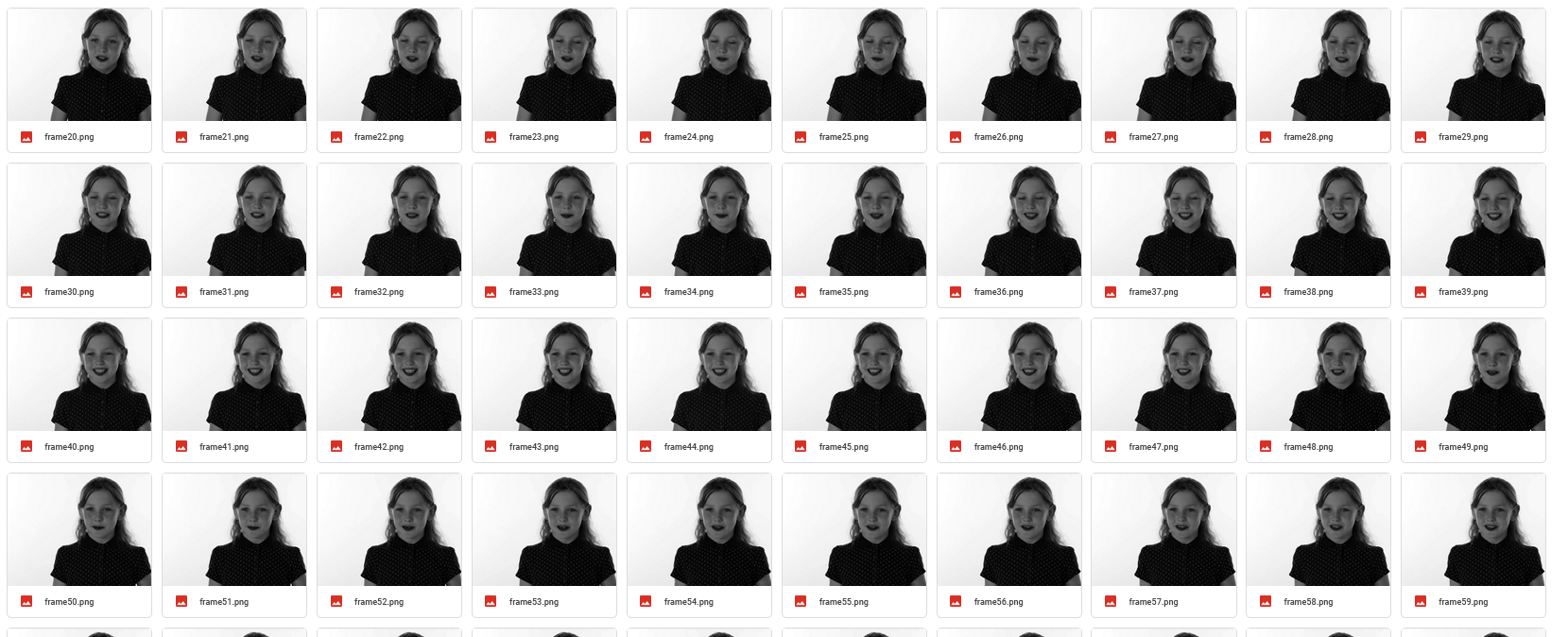


Figure 4.0: Frames taken from the video

100 Frames were taken throughout the whole video. Then The frames were resized into 48x48 grayscale image which was required according to our test Dataset.

Using SoundHound Speech recognition Api support the recognized audio file was later transcript into a text as,

“what’s the best thing about a gypsie on her period when you finger her you get here palm red for free biggest slut in history miss pac-man for twenty-five cents that is a pap smear called a pap smear because girls wouldn't do if it was called scrape the short side and while what do you call a cheap circumcision difference between a walrus and a lesbian.”

After Removing stop words from the raw transcript it changed significantly into,

“ what's best thing gypsie period finger get palm red free biggest slut history miss pac-man twenty five cents pap smear called pap smear girls called scrape short side call cheap circumcision difference walrus lesbian “

Thus, our text is ready for processing. Tokenization was automatically done through the process of Text2Emotion as its own feature.

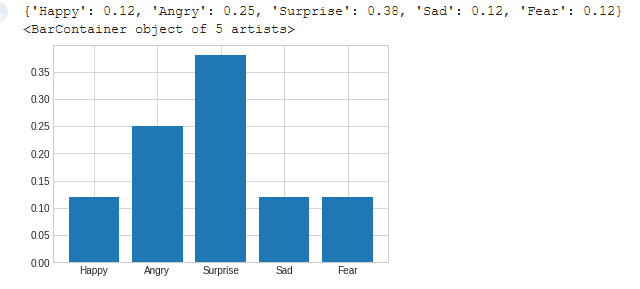


Figure 4.1: Results extracted from the video

The Above image shows the Result Found from the extracted context from the video. On the Other hand, from the video recognition we get,

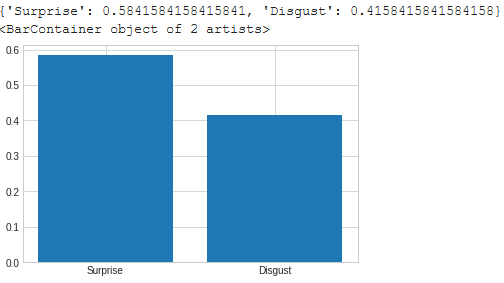


Figure 4.2: Video Recognition

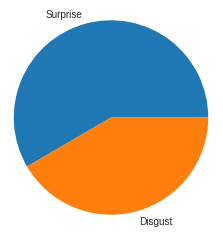
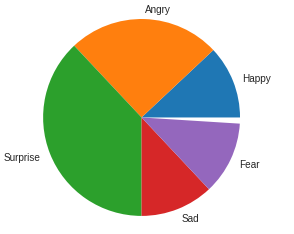
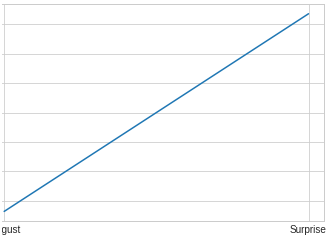
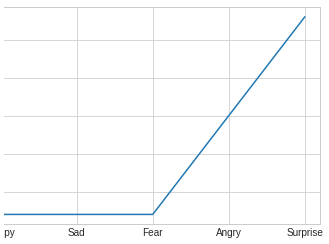
A comparison Pi-Chart of these two recognition is given below,  

Figure 4.3: Pi-Chart

Then we Merged Both result Dictionaries into one final dictionary given priority to the image recognized result. Which gives us a final picture of the recognition.

{'Happy': 0.12, 'Angry': 0.25, 'Surprise': 0.5841584158415841, 'Sad': 0.12, 'Fear': 0.12, 'Disgust': 0.4158415841584158}

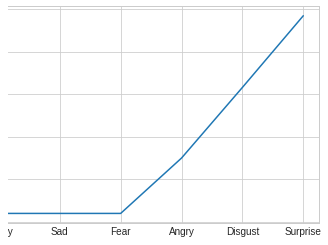


Figure 4.4: Merged Final Result

|  |
| --- |
| **Chapter 5: Future Work** |

Different tests and experiments have been left for the future due to the lack of time (i.e. the experiments with real data are usually very time consuming, sometimes requiring even days to finish a single run). Future work concerns the deeper analysis of particular mechanisms or simply curiosity.

There are some improvements we can do to further improve our system. For our training data set in the future, we can use a custom-made dataset rather than using preexisting data we have taken from others. We can also use the custom model for text recognition. We can also include gesture detection to further increase the accuracy of our model. But the most important aspect of our work is to decrease the cost of implementing the model in practical problems.

|  |
| --- |
| **References** |

[1] Awad, W. A., ELseuofi, S. M. (2011). MACHINE LEARNING METHODS FOR SPAM E-MAIL CLASSIFICATION. International Journal of Computer Science and Information Technology (IJCSIT), 4(5), 2352–2355.

[2] Almeida,tiago. Almeida, Jurandy.Yamakami, Akebo ” Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers” Journal of Internet Services and Applications, Springer London , February 2011 pp.68–73.

[3] Trivedi, Shrwan Kumar. “A Study of Machine Learning Classifiers for Spam Detection.” 2016 4th International Symposium on Computational and Business Intelligence (ISCBI), Sept. 2016, doi:10.1109/ISCBI.2016.7743279.

[4] Barbara Kitchenham, O. Pearl Brereton, David Budgen, Mark-Turner,John Bailey, Stephen Linkman(2009), Systematic literature reviews in software engineering – A systematic literature review, Elsevier.

[5] Yuchun Tang, Sven Krasser, Yuanchen He, Weilai Yang, Dmitri Alperovitch “Support Vector Machines and Random Forests Modeling for Spam Senders Behavior Analysis” IEEE GLOBECOM, 2008

[6] Jane Webster, Richard T. Watson (2002), Analyzing the Past to Prepare for the Future: Writing A Literature Review, MIS Quarterly Vol. 26 No. 2, pp. xiii-xxii.

[7] Shripriya Dongre, Prof. Kamlesh Patidar “E-Mail Spam Classification Using Long Short-Term Memory Method.” International Journal of Scientific Research and Engineering Trends, vol. 5, no. 5, 2019, pp. 1659–1665.

[8] Enrico Blanzieri,Anton Bryl(2009), A survey of learning-based techniques of email spam filtering, Artif Intell Rev, DOI 10.1007/s10462- 009-9109-6.

[9] Asghar, Muhammad Zubair, et al. “Sentence-Level Emotion Detection Framework Using Rule-Based Classification.” Cognitive Computation, vol. 9, no. 6, 2017, pp. 868–894., doi:10.1007/s12559-017-9503-3.

[10] Samson, Andrea C., et al. “Eliciting Positive, Negative and Mixed Emotional States: A Film Library for Affective Scientists.” Cognition and Emotion, vol. 30, no. 5, 2015, pp. 827–856., doi:10.1080/02699931.2015.1031089.

[11] Strapparava, C., Mihalcea, R.: SemEval-2007 task 14: affective text. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pp. 70–74 (2007)

[12] Ekman, P.: An argument for basic emotions. Cogn. Emot. 6(3–4), 169–200 (1992)

[13] International Survey on Emotion Antecedents and Reactions data set. https://www. unige.ch/cisa/index.php/download file/view/395/296/

[14] Das, D., Bandyopadhyay, S.: Sentence level emotion tagging. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–6. IEEE (2009)

[15] Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: Proceedings of the 2008 ACM Symposium on Applied Computing, pp. 1556–1560 (2008)

[16] Francisco, V., Gerv´as, P.: Exploring the compositionality of emotions in text: word emotions, sentence emotions and automated tagging. In: AAAI-06 Workshop on Computational Aesthetics: Artificial Intelligence Approaches to Beauty and Happiness (2006)

[17] Asghar, M.Z., Khan, A., Bibi, A., Kundi, F.M., Ahmad, H.: Sentence-level emotion detection framework using rule-based classification. Cogn. Comput. 9(6), 868–894 (2017)

[18] Shaheen, S., El-Hajj, W., Hajj, H., Elbassuoni, S.: Emotion recognition from text based on automatically generated rules. In: IEEE International Conference on Data Mining Workshop, pp. 383–392 (2014)

[19] www.theneweconomy.com/technology/the-problem-with-emotion-detection-technology

[20] X. U. Feng and J.-P. Zhang, ‘‘Facial microexpression recognition: A survey,’’ Acta Automatica Sinica, vol. 43, no. 3, pp. 333–348, 2017.

[21] M. S. Özerdem and H. Polat, ‘‘Emotion recognition based on EEG features in movie clips with channel selection,’’ Brain Inf., vol. 4, no. 4, pp. 241–252, 2017.

[22] S. K. A. Kamarol, M. H. Jaward, H. Kälviäinen, J. Parkkinen, and R. Parthiban, ‘‘Joint facial expression recognition and intensity estimation based on weighted votes of image sequences,’’ Pattern Recognit. Lett., vol. 92, pp. 25–32, Jun. 2017.

[23] Hongli Zhang , Alireza Jolfaei , and Mamoun Alazab, “A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing,” IEEE ACCESS, Nov, 2019. 2949741

[24] H. Ma and T. Celik, ‘‘FER-Net: Facial expression recognition using densely connected convolutional network,’’ Electron. Lett., vol. 55, no. 4, pp. 184–186, Feb. 2019.

[25] A. V. Savchenko, ‘‘Deep neural networks and maximum likelihood search for approximate nearest neighbor in video-based image recognition,’’ Opt. Memory Neural Netw., vol. 26, no. 2, pp. 129–136, Apr. 2017.