Multi-modal Emotion Recognition Using Deep Learning Architectures

Author

Iram Hina

Regn Number

FALL 2016-MS-16(CSE) 00000171137

MS-16 (CSE)

Thesis Supervisor:

Dr. Arslan Shaukat

DEPARTMENT OF SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL AND MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

JULY, 2020

Multi-modal Emotion Recognition Using Deep Learning Architectures

Author

Iram Hina

FALL 2016-MS-16(CSE) 00000171137

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Software Engineering

Thesis Supervisor:

Dr. Arslan Shaukat

Thesis Supervisor's Signature: 

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY, ISLAMABAD

JULY, 2020

DECLARATION

I certify that this research work titled "*Multi-modal Emotion Recognition Using Deep Learning Architectures*" is my own work under the supervision of Dr. Arslan Shaukat. This work has not been presented elsewhere for assessment. The material that has been used from other sources, it has been properly acknowledged / referred.

_____

Signature of Student

Iram Hina

FALL 2016-MS-16(CSE) 00000171137

_____

Signature of Supervisor

LANGUAGE CORRECTNESS CERTIFICATE

This thesis is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the University for MS thesis work.

_____

Signature of Student

Iram Hina

FALL 2016-MS-16(CSE) 00000171137

_____

Signature of Supervisor

COPYRIGHT STATEMENT

ACKNOWLDGEMENTS

I am highly thankful to ALLAH Almighty for his blessings and constant help throughout my thesis. Indeed, this would not have been possible without his bountiful and gracious help in each and every step. I am thankful to HIM for putting me on a path where people helped me graciously to achieve my goal. Indeed, none is worthy of praise but ALLAH Almighty.

I am profusely grateful to my beloved parents for their constant love, support, prayers and sacrifices. I would like to pay my special thanks to my beloved father **Muhammad Taqdees,** my mother **Shehnaz Akhtar** and my brother **Muammad Musadiq**, for encouraging me to avail best opportunities in my life. I am also thankful to my sisters and family for all the support and prayers throughout my time of research.

I would also like to express my gratitude to my supervisor **Dr. Arslan Shaukat** and my co-supervisor **Dr. Usman Akram** for their constant motivation, patience, enthusiasm, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my MS study.

I would also like to pay special thanks to my Guidance Committee Members **Dr. Farhan Hussain and Dr. Sajid Gul Khuwaja**. Their recommendations are very valued for improvement of the work.

I would also like to thank **Muhammad Bani Asif** for his motivation and moral support throughout this journey, my class fellows, my friends and my seniors for their support and cooperation. Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

Thanks for all your encouragement!

Dedicated to my exceptional parents: ***Muhammad Taqdees & Shehnaz Akhtar,*** *adored siblings and wonderful friends whose tremendous support and cooperation led me to this accomplishment*

ABSTRACT

Emotions are essential part of immaculate communication. Emotions play a vital role in decision making, behavior learning and communication activities in daily routine. Speech communication and facial expressions are considered to be the root of information. The purpose of this research is to design an automated system which can recognize six basic emotions of human namely anger, disgust, fear, happiness, sadness and surprise for effective communication between humans and computers. In proposed method audio-visual features from videos with emotions have been extracted separately. A sequential deep convolution neural network (CNN) has been used along with Recurrent Neural Network (RNN) to classify these emotions. From audio, features like MFCC have been extracted and passed to CNN for audio classification. In comparison fine tuning has been performed on pre-trained AlexNet deep CNN having mel-spectrogram as input. Features extracted from fine-tuning of AlexNet give better recognition rates on audio data. On the other hand, visual features have been extracted from video frames using CNN and then fed to the RNN using LSTM layer to handle the temporal nature of experimental data. Multimodal emotion recognition has been performed by fusing audio and visual modalities together though decision level and score level fusion. SVM, random forest, K-NN and logistic regression classifiers were used to classify emotions from fused audio-visual data. Experiments have been performed on two audio-visual databases namely RML and BAUM-1s. RML contains 720 video samples recorded by 8 actors and BAUM-1s contains 544 video samples recorded by 31 actors belonging to different ethnic and cultural background. Leave-One-Speaker-Out (LOSO) and Leave-One-Speaker-Group-Out (LOSGO) cross validation techniques are used for evaluation of our model on RML and BAUM-1s respectively. Competitive recognition rates are achieved on both datasets i.e. 61.68% on BAUM-1s and 79.51% on RML. The recognition rate on BAUM-1s dataset is 61.68 % which is an improvement over previous state of art results by 1.19%.

Keywords: Audio-Visual Emotion Recognition, Multi-modal, Deep Convolution Neural Network, Deep Learning, Recurrent Neural Network, Long Short Term Memory, CNN-LSTM

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

CHAPTER1 : INTRODUCTION

This chapter provides thorough introduction of this research including motivation, problem statement, applications, research contribution and thesis organization.

Technology has been spinning the wheel of revolution since the establishment of computer science. With this revolution of technology there is always a room for improvement and this never ending expedition has led humans to automate their tasks and productivity. The capability of a system performing tasks which requires human intelligence is just marvelous. With the increase in modern technologies artificial intelligence has grown itself into two major branches: psychological research and physiological research which is based on human brain thoughts and activities. The purpose of Strong AI phenomenon is to reach to a point in artificial intelligence where a machine's rational capability will be equal to a human. This concept leads to learning of human behavior as well as to understand leaning capabilities of human.

1.1 Motivation:

We human are created as emotional beings and due to this characteristic humans cannot convey their message properly without having expressions and emotions during a conversation. To properly communicate our intentions and feelings to others we need emotions. Likewise, our communications with machines can be improved momentously if they will be able to interpret the message properly which is only possible if machine will be able to differentiate among basic emotional states of the sender of the message. Audio-Visual emotion recognition plays a vital role in the improvement of human computer interaction. Major applications of this area lie in the field of pattern recognition, robotics, medical diagnosis, information processing and mobile computing. With the growth in technology and automation we need machines that can carry out our tasks on daily basis. For that purpose we need machines to be intelligent to an extent where they can learn, adapt and understand human behavior and act accordingly.

1.2 Problem Statement:

For humans, it might seem very easy to distinguish among different emotional states because we are designed this way but for a machine it is a quite challenging task. Special learning techniques and different training algorithms are required to accomplish this task. A lot of research has been

performed in this area. Correct recognition and distinction of different emotional states using different machine learning algorithms along with computer vision techniques is a complex job to do and it depends on several factors that if neglected, can have a negative impact on performance of the system. Major factors that can affect the performance of the system include noise, occlusion, voice quality, illumination changes and most significantly variety of facial textures due to different regions and disparate emotional expressions between various people.

1.3 Research Contribution:

Detailed set of contributions of the proposed approach are as follows:

- Review and comparison of recent and most impactful advancements in automatic emotion recognition systems.
- Extraction of most discriminative audio features from video clips.
- Designing of unique deep convolutional neural network that can extract visual features from videos and classify accordingly.
- Extraction of visual features using already designed powerful deep convolutional neural network for fine tuning and comparison with our proposed network architecture.

1.4 Thesis Organization:

Organization of thesis can be viewed in figure 1.1.

Chapter 2 covers the importance of emotions and their categorization techniques as well as basic emotion recognition techniques and neural networks.

Chapter 3 covers the review of literature which includes the previous contributions in the domain of emotions recognition. Literature review has been performed in two steps: Firstly, research articles are evaluated to explore different contributions in the domain of emotion recognition. Secondly, multiple research articles are studied to analyze different implementation techniques for emotion recognition. Finally, research gap has been identified from literature review.

Chapter 4 presents the proposed methodology and feature extraction strategy used for audio feature extraction.

Chapter 5 covers validation techniques used for evaluation of proposed methodology

Chapter 6 discusses limitations, conclusion and future work of this research.

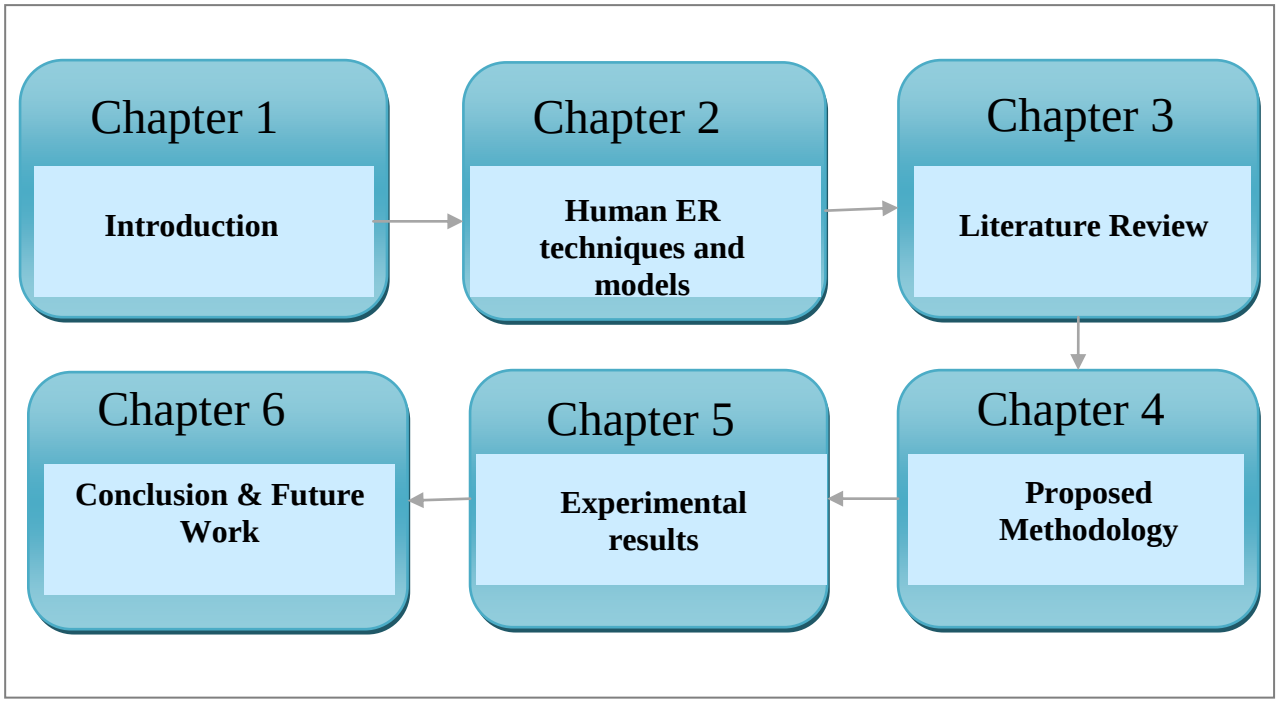

Figure 1-1: Thesis Organization

Chapter 2

Human Emotion Recognition Techniques and Models

CHAPTER2 : HUMAN EMOTION RECOGNITION, TECHNIQUES AND MODELS

Relevant concepts of this research are covered in this chapter. Sequential review of emotion classification, feature extraction, machine learning, affective computing and other relevant concepts have been performed. This chapter also includes discussion about related subjects.

2.1 Affective Computing:

Affective Computing is the field of study which focuses on development of such a system or device that can perform recognition and interpretation of human emotions and then process that information to arouse human emotions in machines [1]. According to Rosalind Picard [2], communication, decision making and learning are core phenomenon that are based on emotions.

Human being is considered to be a rational decision maker which is not possible without having emotions. Although emotions play a vital role in decision making but according to some researchers decision could be highly biased based on emotions [3]. Empathy is a basic trait of human which allows other humans to understand and feel what other person is feeling and what that person is going through [4].

The basic motivation behind affective computing is the ability of stimulation of empathy in computing systems. In which a system can understand various emotional states of human and adapting their behavior to provide suitable response to those emotions [5]. With the growing technology we want computers to be smart enough to be polite and more socially responsible without processing irrelevant information. For this purpose emotional states of humans need to be critically analyzed.

Recognition and interpretation of emotions involves meaningful patterns extraction from the collected data. Different machine learning techniques are used to process given information and extract meaningful patterns such as speech recognition, facial expression recognition, natural

language processing. This research work focuses on two modalities speech recognition and facial emotion recognition.

2.2 Emotion Classification:

Emotion classification is based on two main perspectives. First perspective states that emotional states are independent of each other and are therefore distinct and discrete. There is not interconnection among any of emotional state. Whereas, second perspective states that basic emotional states are interlinked and can be grouped together on dimensional basis.

2.1.1 Basic Emotional States:

It is quite easy for humans to understand emotional states and interpret message behind those emotions because humans are created this way. This ability of emotion recognition has developed a categorization of basic emotional states which are constant among all the people. Numerous researches have been conducted to find out basic emotional states [6]-[8]. According to Ekman rt al. [9] emotions can be classified into six basic universal categories namely anger, disgust, fear, happiness, sadness, surprise. These emotional states are distinct and have their own particular characteristics. He also stated that these emotions can also be considered as a category instead of separate emotional states because of their representation in terms of degrees.

2.2.1 Multi-dimensional Representation:

Instead of dividing emotions into discrete categories, researchers have proposed dimensional representation for emotional states which allows the visual display and better understanding of emotion distance between different experiences. Thayer et al. [10] proposed a multi-dimensional representation to state emotional experiences where each emotion state is mapped on to a two dimensional plane. In his proposed representation, each emotion is to be weighed in terms of its activation-valence mapped in two dimensions. Activation/arousal depicts the degree of excitation of each emotion whereas valence represents how positive or negative the emotional experience is. Figure 2.1 shows the two dimensional activation valence diagram.

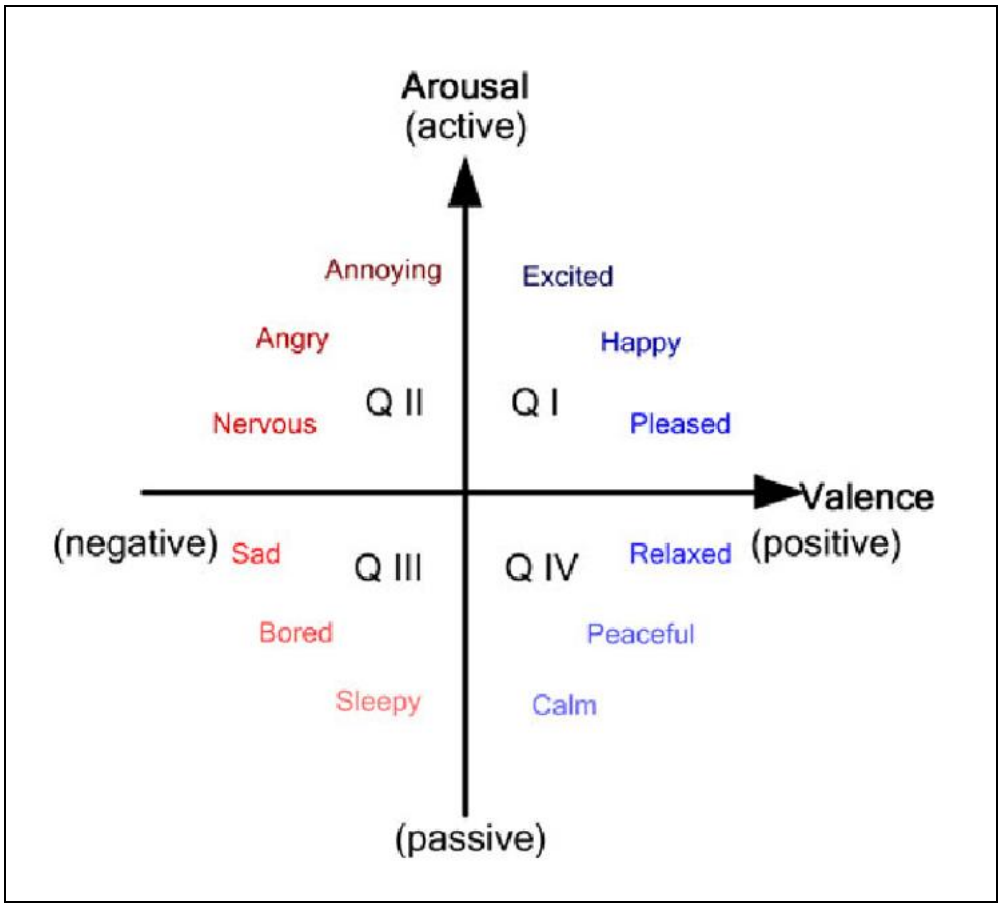

Figure 2-1: 2D emotional Plane [10], [11]

2.3 Typical Audio-Visual Emotion Recognition System:

After the complete understanding of emotions, their recognition and their classification process researches started to wonder about how the emotions can be learned and predicted automatically. Normal representation of emotions can be predicted via facial expressions, speech and different postures [11]. Numerous researches have been conducted on these facets taking one or more modalities into consideration. Our research focuses on two modalities: audio and visual content from videos to distinguish among different emotions. A basic audio-visual emotion recognition system is a combination of three steps: First step is extraction of audio and images from video data, Second steps leads to feature extraction from audio and images and the third and final step focuses on classification of distinct emotions based on extracted features. Following block

diagram 2-2 show all three steps of a basic audio-visual emotion recognition system. The details of system is explained below.

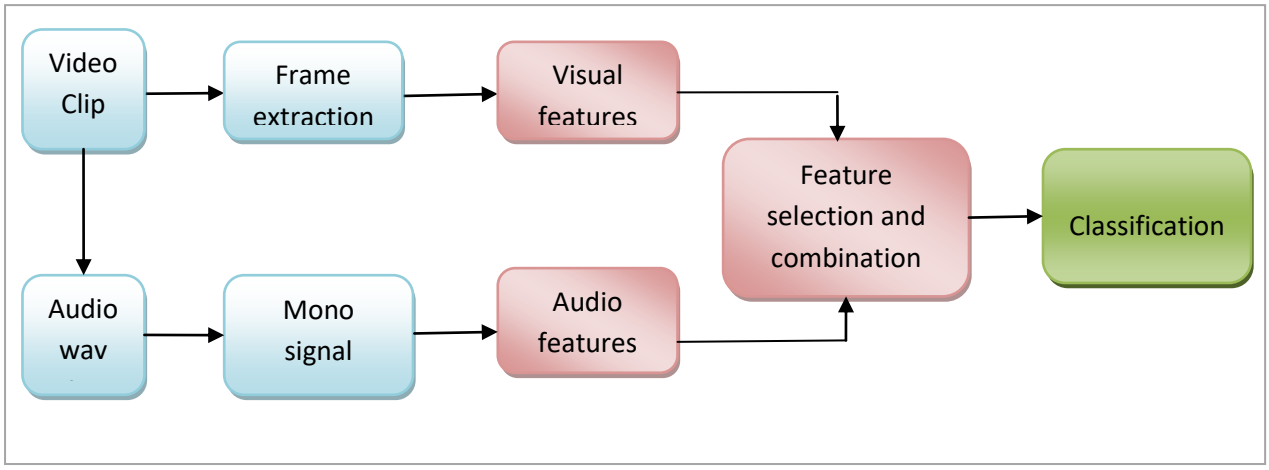

Figure 2-2: Typical Audio-visual Emotion Recognition System

2.3.1 Audio-visual Corpus:

When we talk about experimentation the first thing we are going to need is audio-visual dataset to conduct experiments. There are many datasets available for experimentation purpose. Researches use those datasets to conduct experiments or create their own. For audio-visual analysis there are three major types of datasets that are available.

- **Spontaneous:** Spontaneous dataset contains natural emotional observations and there could be multiple sources from which data is being recorded. Emotional response to any kind of interaction is totally natural. Spontaneous datasets are considered to be very challenging because of variation in speaker's angles and positions. Change in position makes it harder to label the data as well as to extract visual features. Classification based upon audio features also become less reliable because of natural responses. Examples of such kind of datasets are RECOLA, BAUM-1s, BAUM-2s and SEWA DB etc.

- **Induced:** This type of datasets contains responses which are recorded by induction of emotions. In this category emotions are induced in speakers by providing them some kind of stimulus for example showing them any kind of emotional or horror video clips. SAVEE dataset is one of the examples of this category.

- **Posed:** In posed/acted datasets speakers are provided with a script on which they act and their responses are recorded accordingly. The observations are according to the script which makes it easier to label the data and extract visual features as speakers are facing the camera directly. RML, eNTERFACE'05 are examples of acted datasets.

We have chosen RML as acted dataset and BAUM-1s as spontaneous dataset for experimentation of our emotion recognition system.

2.3.2 Feature Extraction:

In this phase features needs to be extracted from video data. However, features are extracted separately from audio and visual content. For this purpose, audio content has been separated from videos following some pre-processing then eventually feature extraction has been performed.

1. **Audio Feature Extraction:**

Audio features are extracted from audio content of the videos. Audio data is then processed to extract features which are further used for classification of the emotions based on speech data. Ample amount of research has been performed on speech emotion recognition in past few years. Researchers have used different techniques to extract features that lead to better emotion recognition system. Some common audio features are: prosody features which includes pitch, loudness and duration of speech and present difference in the prosodic patterns depending on speaker's speaking rate, pitch range, phrasing and inflection. Then are spectral features which are formulated from short term audio signals. Some common examples of spectral features are Mel Frequency Cepstral coefficients (MFCC), which mimic the human auditory response. Other common spectral features are Linear Predictive Cepstral Coefficients (LPCC), Power Spectral Density (PSD), RASTA-PLP and Formant frequencies which can be observed as choral tract resonances.

Both prosodic and spectral features are useful for speech emotion recognition as they are extracted from audio content. However, some researchers also claims MFCC features to be the better performers than others [12].

2. **Visual Features:**

After extraction of audio features from speech content, next step is to extract visual features from videos by detection of face through several frames in the video and from those frames useful information as features is extracted for classification. There are two types of features that can be extracted 1: geometric features and 2: appearance features [13]. The geometric features include information about specific face portions such as eyes, mouth, eyebrows and their corner points. These features are often used by researchers in their experiments, however, geometric feature based methods are usually difficult to implement because it requires precise and dependable facial feature detection which can create problems in many situations. The appearance features on the other hand deal with the whole face or a certain region on it, without concentrating on a specific portion of the face.

As with the growing industry there is a revolution in this field as well. Instead of using hand crafted features learning features are being used now. Hand crafted features are those features which are calculated manually by researchers. In this approach first we define a set of features and then extract them. Whereas, learned features are those feature which are attained from an affective machine learning algorithm in which a machine learning model train itself to extract useful features and decides on its own that which features are best to extract [14]. Hand crafted feature extraction was used traditionally for emotion classification however, with the birth of deep learning things have changed [15]. That's why we have also selected deep learning models to extract features and classify them accordingly for our system. A common algorithm used for hand crafted features is local binary pattern (LBP) whereas; learned features are commonly extracted using convolutional neural network (CNN) [16].

2.3.3 Classification:

The final step in emotion recognition system is classification which has separately performed for both modalities audio and visual content. We have used deep learning models for both modalities to classify emotions.

2.4 Machine Learning:

Machine learning and artificial intelligence are interlinked and are highly correlated. To create an intelligent system these both technologies are used in connection with each other. Artificial intelligence is study and development of such a machine that is capable of thinking and behaving like human whereas, machine learning is considered to be a subset of artificial intelligence in which system does not need to be programmed explicitly but it learns from input data. Arther Samuel was the first person to describe the term machine learning in 1959. In a traditional software development computer is destined to execute some set of instructions given by the developer of the program whereas, machine learning differs in this context because it focuses on discovering an algorithm that can develop or improve itself on its own instead of being explicitly programmed. Machine learning mainly emphases on decision making or predictions from the given data that is why is also termed as predictive analytics [17].

There are three main types of machine learning algorithms named as: supervised learning, unsupervised learning and reinforcement learning. In supervised learning algorithm generates a model which contains inputs and desired outputs. Inputs given to the system are labeled in this case belonging to particular categories. Regression and classification mainly lies under supervised learning. In contrary unsupervised learning contains only input data without having categorization labels therefore, in this case algorithms learns from unlabeled data and find sequence and structure of data to take decisions. Cluster analysis is an example of unsupervised learning. Third category reinforcement learning focuses on the actions and behavior of an agent in different environments to achieve a goal, based upon which the agent gets rewarded or punished. Chess game is an example of reinforcement learning.

This research falls under supervised learning because the data being used for experiments is labeled. Label here corresponds to a unique emotion.

2.5 Artificial Neural Networks:

An artificial neural network (ANN) is a model that process information in the same way a brain does. ANNs are designed on the brain structure to make machine more reasonable and think like human. An ANN is an endeavor to work the same way as a network of neurons works in the brain. In early 1940s scientist were working on the concept of simulating brain structure but

Warren McCulloch and Walter Pitts were the first to present artificial neural network (ANN) in 1943 with a computational model [18]. Our brain has number of neurons which are being connected with one another. There are sensory organs for various inputs which are being fed to the neurons and according to the different input and issues neurons can send message to other neurons. ANN works on the same principle. Just like our brain an ANN is comprised of multiple nodes (neurons in brain). These nodes are connected through links. An ANN use different layers to process the data that it is given.

There are three things that make up an artificial neural network: Input layer: consists of multiple nodes responsible for receiving input data. This input data is further used for process and learning. Hidden layer: this input data then passed to a hidden layer whose job is to transform data into something that is useful for output layer. Weights are used to transform input data. In most of the networks nodes of one layer are fully connected to the nodes of next layer and the connections between layers are weighted. The more the value of weight the greater will be the impact on the nodes. Figure 2.3 shows a simple ANN structure.

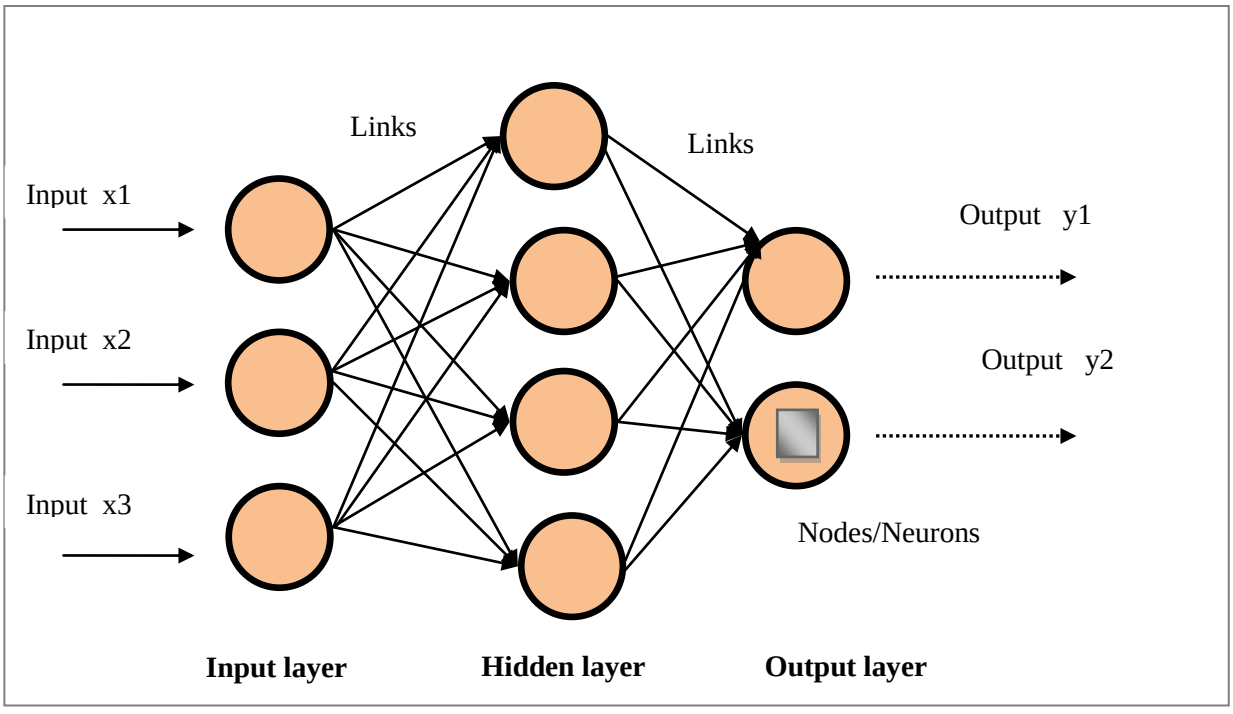

Figure 2-3: Artificial Neural Network Structure [18]

The output of layer on the left becomes input of next layer with the help of weighted sum. And in the end the output of hidden layer will become input for the output layer.

Perceptron was the first implementation of a simplest neural network. Researchers took ten years to implement this first neural network [19]. And for the first time an artificial neural network was able to learn by means of supervised learning. Figure 2.4 shows the architecture of a perceptron.

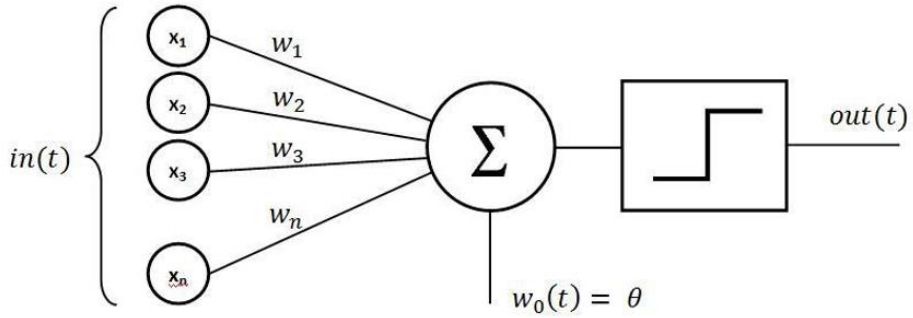

Figure 2-4: Perceptron Architecture [19]

This simplest architecture is the base of most of ANN models. The inputs x_1, x_2 ,…, x_n represents the input nodes and layer containing them is the input layer. W_n is the corresponding weight for each input. A weighted sum is performed on the node. In order for nodes to perform as a linear function a bias term is added (θ). The bias term is a constant term which is added to adjust the output with weighted sum. It also helps the model to fit best for the data that is being used for processing.

$$y = f(t) = \sum_{i=1}^{n} X_i * W_i + \Theta \tag{2.1}$$

The result of f(t) becomes the input of activation function. As the perceptron is a binary classifier so a binary step function is used as an activation function which states that the output will be 0 or 1.

$$output = \begin{cases} 0, & y < 0 \\ 1, & y \geq 0 \end{cases} \tag{2.2}$$

At the end the predictions are measured using the real values. The weights on the first layer are updated from the error came out at the output layer. Process of updating weights from error signal is performed by back propagation.

ANNs were quite popular in early 1960s but in 1969 Minsky and Papert discussed the limitation and drawbacks of perceptron in their research and stated that the performance of neural network could not be improved by making the architecture more complex. The limitations of ANN were major issue back then but now with the revolution in technology those obstacles have been conquered successfully. Slow but continuous progress for improvement of ANN was in action since 1960s but the real turn came in early 2010s. One main reason for this exponential growth was invention of new hardware and computational devices which can process tons of data in a blink. The rise of Big Data has taken machine learning to a new level. Thus, all these new innovations and discoveries gave birth to deep learning which replaced ANNs eventually.

2.6 Deep Learning:

Deep learning is an extensive form of machine learning which is based on ANN architecture with a difference of having multiple layers to extract information from input data [20]. Data here is transformed through multiple layers. The term “deep” is used to refer to the number of hidden layers used in the network. Explicit feature extraction using hand crafted techniques is not required in deep learning. Deep learning models are capable of learning features on their own. For instance in image processing, lower layers in deep neural network may be used for identification of edges and corners and higher layer may work on extracting information about human, faces and letters.

At each layer input data becomes more abstract and composed. For instance, in case of image processing, the input layer is given a raw image in the form of a matrix having pixels information. The first layer might calculate edges and corner, the task performed by second layer could be arrangement of the detected edges and corners whereas, third layer could be identifying nose and eyes portions and the job of fourth layer will be recognition of human face in the

image. Hence, deep learning is capable of learning features and deciding that which features are to be passed to which layer [21].

The theoretical concept of deep learning was introduced by Rina Dechter in 1986 but the first fully functional deep neural network with standard back propagation was used for handwritten zip codes by Yann LeCun et al. in 1989, whose training took three days to get completed [22]. The rise of deep learning in the computer industry started in early 2000s when Convolutional Neural Networks (CNNs) were used to process checks transcribed in US. CNNs have been used for years but didn't get significance till 2011. In 2012 Ciresan et al. [23] presented CNNs along with max pooling trained on GPU with a dramatic improvement in many computer vision benchmarks. Similary Krizhevsky et al. [24] presented better results on imageNet classification using CNNs.

In 2003, a research on speech recognizers was conducted in which Long Short Term Memory (LSTM) network was used which outperformed the traditional approaches used for speech recognition system [25]. LSTM is a special type of artificial neural network also known as Recurrent Neural Network (RNN) which keeps the memory of events occurred thousands of distinct time steps before the current sequence. In 2015, RNN having LSTM layer was used for Google's speech recognition system and performance rate was increased by 49% [26]. Both CNN and RNN have been used in our proposed model in combination for emotion recognition.

2.7 Convolutional Neural Networks:

Convolutional neural networks are considered to be more reliable and stable version of perceptrons having multiple layers. Multilayer perceptrons are also referred to as fully connected networks in which each node (neuron) of one layer is linked to all the nodes of the next layer. These fully connected networks prevent overfitting of data. Overfitting occurs when a model fails to generalize i.e. a model start learning details along with the noise to an extent that it becomes perfect for training data but shows negative impact on the unseen data that we generally call as testing data.

The development of the convolutional neural networks is related to the study of visual cortex. Hubel and Wiesel conducted a study in 1968. Visual cortexes of monkeys and cats were

presented in this study stating that they their cortexes consists of neurons which respond to tiny sections of visual fields individually. Their study also focused on two types of visual cells present in brain: simple cells and complex cells. Simple cells works in particular receptive field and covers edge like shapes with specific orientation whereas, complex cells work on a broader receptive field irrespective of arrangement and orientation of edges [27]. Receptive field is referred to a region which is responsible for elicitation of responses occurs due to any kind to stimulus.

Neocognitron was one of the early implementations based on the ideas of Hubel and Wiesel. Kunihiko Fukushima developed a neural network model known as Neocognitron [27] in 1980. Two basic layers were defined in Neoconitron: First was Convolutional layer which was formulated by simple cells represented by units and second was downsampling layer which was formulated by complex cells. One of the greatest achievements of Neocognitron includes the implementation of the local invariant property. Moreover, there is one to one output mapping in it. One and only one specific pattern is mapped by a complex cell.

The learning process of Neocognitron is the main drawback. In the past, there was no method to measure errors. In 1985, the backpropagation modern form [28] derived by Finish Mathematician Seppo Linnainmaa in 1970 [29] [30] also got applied in ANN. But its use was limited at that time and few applications were developed using backpropagation [28]. The use of backpropagation was introduced into ANN in 1985 by Hinton, Rumelhart, and Williams [31]. The gradient of the error is measures by backpropagation with respect to the weights on the units. Gradient changes with an increase or decrease in the value of weights. After that, the gradient will be used to find weights thereby minimizing the network error. When using backpropagation, the network is able to auto-tune its parameters with some optimizer such as Gradient Descent (GD) which is a first-order optimization algorithm.

Convolutional Neural Networks are most commonly used for analysis of visual data. They are also called shift invariant because their architecture is based upon weight sharing and translation invariance. Translation invariance is a processing of moving an object from one place to other without rotation.

The key components of a CNN are; Convolutional layer which performs Convolutional operation, Weight sharing, Receptive Fields, Pooling layer responsible for pooling operation/spatial sub sampling, Dropout, ReLU layer works as activation, Stochastic Gradient Descent, Classification with fully connected layer and loss layer.

2.7.1 Convolutional Layer:

Main building block of a ConvNet is the convolutional layer, which is responsible for convolutional operation. Convolution is a mathematical concept which is combination of two functions. This mathematical function works as a filter to avoid passing unnecessary information to the next layers of the network and it requires two elements; input data (an image in the form of a matrix having pixel values) and a filter/kernel. Main purpose of convolutional operation is extraction of high level features for instance edges from the given input image.

The input of a convolutional layer is a tensor having a shape as (number of samples) x (sample width) x (sample height) x (sample depth). Here sample could be an image and depth refers to number of channels either its colored or greyscale. In case of a colored image the depth is 3 which represent 3 color channels RGB (Red, Green and Blue values) for each pixel in image and in case of a greyscale image the depth will be 1 representing 1 channel. In further layers of a CNN number of channels could be more than three representing abstract form of colors acting like RGB channel and every channel depicts some useful information about the transformed image.

The resultant of a convolutional operation is feature map. After the operation of convolution we get a feature map having a shape (number of samples) x (feature map width) x (feature height) x (feature map channels)

A convolutional layer must have following three attributes:

- A convolutional filter of shape (filter width) x (filter height). For instance 3x3x1 is for greyscale image where 3x3 represents width and height of the kernel and 1 depicts number of channels. Whereas, in case of an RGB image it will be 3x3x3.
- Number of channels for input and output.

- Depth (number of channels) of a convolutional kernel must equate the depth (number of channels) of input image.

Figure 2.5 shows the movement of the filter over an image with a depth of three.

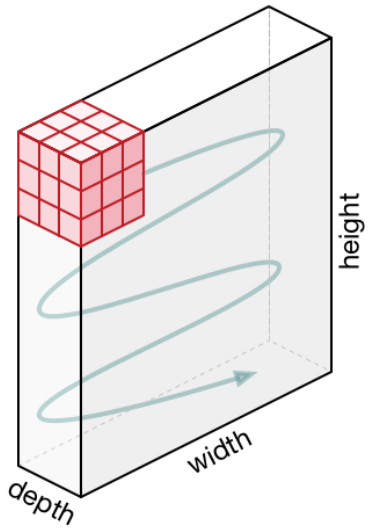

Figure 2-5: Movement of Kernel [27]

Number of feature maps turns out to be the same as number of kernels used for convolution on the input data. These feature maps are independent of each other and helps to learn new features. Figure 2.6 shows a graphical representation of convolution for edge detection.

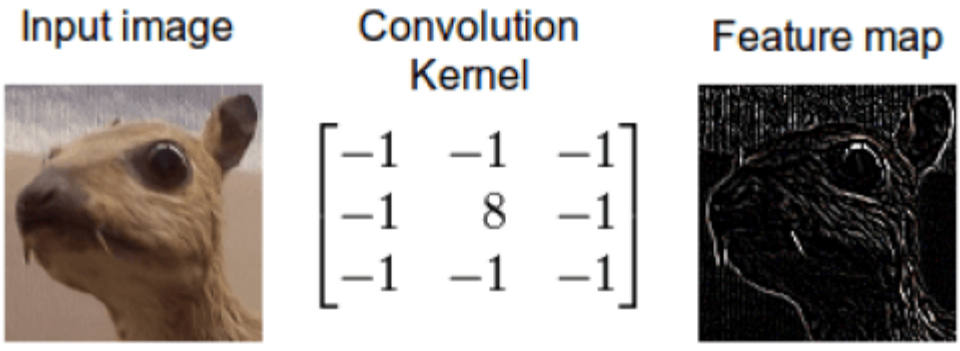

Figure 2-6: Convolution using edge detector filter [31]

In case of a colored image having multiple channels such as RGB the kernel applied will be of similar depth. Convolution is performed across all three channels and then the results are being added along with a bias term to provide compacted convolved feature map having one channel. The resultant of the convolution can be obtained with two types; First type refers to reduction in dimensions of feature map in comparison to the input; this can be done using valid padding. In second type dimensions of the feature map turns out to be same as input, this is done using same padding. Figure 2.7 shows a same padding operation

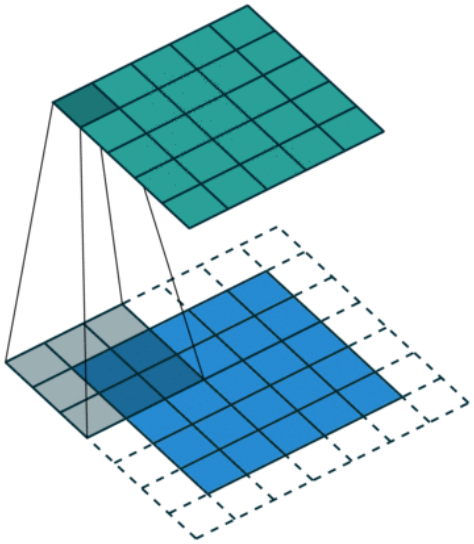

Figure 2-7: Image padded with 0s to get feature map of same dimension [27]

If we take a look at above figure it's an image of 5x5x1 which is being converted into 6x6x1with zero padding and after applying 3x3x1 filter we'll get a feature map of 5x5x1. On the contrary, in valid padding the image is not padded with 0s and after applying 3x3x1 filter to 5x5x1 input image we get a feature map of 3x3x1.

2.7.2 Weights:

Weights in a neural network are used to decrease the error. Weight is a numerical value which is updated through back propagation helps to lessen the error. Regardless of its position, intuition says that a detected feature is always meaningful. High dimensional input's translationally-invariant structure is exploited by weight sharing. Mean to say, the image of a cat is not

recognizable due to cat's position. Another example is; the position of a noun should not change the meaning of a sentence.

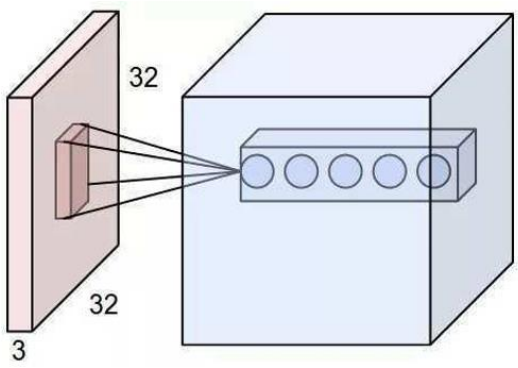

Figure 2-8: Local receptive field of size 5x5x3 for a typical CIFAR-10 image of 32x32x3[31]

A plane is generated after the convolution operation. This place is normally composed of the results of applying the same filter through the entire input. This plane is named feature map [31]. A convolutional operation with a kernel results in a feature map. Different weights initialize kernels to perceive different features. This way, the feature stays for the whole feature map having irrelevant position with respect to the network. For a given filter, weight stays same during convolution.

There is a set of feature maps in a convolutional layer that extract different features at each input location. Convolution operation means the process where input is scanned and unit state is stored on the feature map.

2.7.3 Local Receptive Field:

In real time applications we normally deal with high dimension inputs. In that case it is practically impossible to connect neurons of hidden layer to all the neurons of previous layer. For this purpose only a small region is connected to the neuron, that small region is called receptive field of that particular neuron. The receptive field is always equal to the filter size applied to the input volume. We can also say that receptive field in a CNN is the area of the input which is perceptible to a single filter/kernel at a time. A kernel size, local receptive field and filter size are

similar terms. In figure 2.8 the smaller red area to which lines are connected represents the local receptive field for the neuron that the lines are attached to.

By taking a look at figure 2.8 we can see that the input volume size is 32 x 32 x 3. If the filter size applied to this input is 5 x 5, then every neuron in convolutional layer will be connected to a region of 5 x 5 x 3 making a total of 5*5*3=75 connections +1 bias term. The connectivity of receptive field is local as per space i.e. 5 x 5 but covering the full input depth that is 3. In general it is to be said that if we have three dimensions height, width and depth of an image than local receptive field will only be applicable to height and width and it will not affect depth of the image. Kernel will be focusing on height and width of the input image where as it will be along the full depth. In case of a colored image it will be along 3 axis RGB representing red, green and blue channel. Figure 2.8 represents how a neuron is connected to a feature map of 5x5x3. This process works for an entire image. After this process, the height and width of the input will decrease. This does not happen in case of input depth dimension.

When the convolution operation ends, the filters applied to the input shows the depth dimension. Different features in the image can be captured using a set of filters. Different weights initialize a filter. For a given filter, weights remain same during convolution. The main focus of such operations is to learn useful feature. For optimization of these layers weights are updated using stochastic gradient descent via back propagation. Local receptive field for a fully connected layer is the whole preceding layer.

2.7.4 Pooling Layer:

The most important and common part of a convolutional neural network after convolutional layer is pooling layer, which performs pooling operation. Pooling is responsible for down sampling of the data in a non-linear manner. Convolved feature is reduced in dimensions with respect to spatial size for making data processing easy by decreasing computational complexity and power. Dimensionality is reduced in such a way that outputs of a group of neurons of preceding layer are combined into a single unit for the successive layer. Global pooling works on the neurons of entire convolutional layer whereas, local pooling works on combining the outputs of small groups, usually 2x2 [32]. Furthermore, pooling helps to extract high level features, which are invariant in position and rotation.

Pooling can be performed using two types; Max pooling and Average pooling. Max pooling yields the highest value of the area of the image that a kernel is covering. Whereas, average pooling calculates the mean of all values that are covered by the filter. Pooling is performed along two dimensions of the width and height whereas, depth remains same. Figure 2.9 depicts the types of pooling.

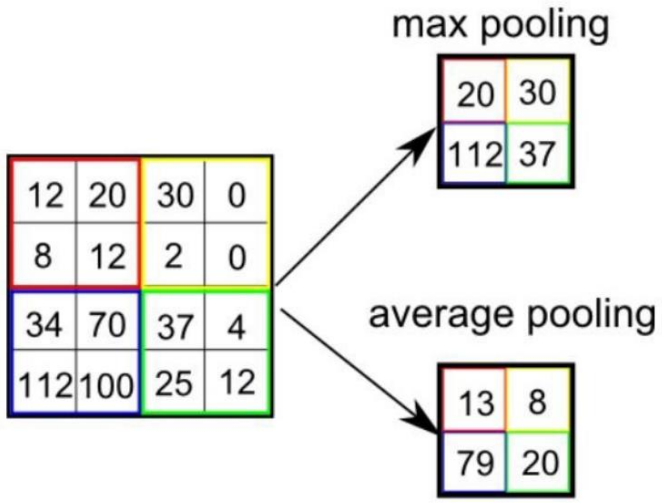

Figure 2-9: Pooling Operations [32]

Max pooling is considered to be far better than average pooling because average pooling only works for dimensionality reduction whereas, max pooling also helps in noise reductions along with reduction in dimensionality. Both pooling and convolutional layers combines to extract dominant features and these layers can be increased in number to get more detailed features with an increase in computational complexity. Pooling also plays an important role in reducing overfitting. At this point, CNN model is capable of learning features on its own and extracting all the useful information required.

2.7.5 Dropout:

By using dropout function, the impact of units with a strong activation can be reduced to a minimum value. To enable other units to learn features automatically, dropout method

shutdowns units during training. All units need specific level of independence to reduce strong unit bias. This leads to better generalization and strong regularization.

2.7.6 Stochastic Gradient Descent:

Gradient Descent refers to an algorithm that is used to find out the updated values for coefficients (parameters) to reduce the value of cost function. This algorithm is mainly used when values of coefficients cannot be determined using linear algebra so the algorithm helps to search the values using optimization. The term stochastic refers to a single training sample. The major difference between gradient descent and stochastic gradient is that gradient descent picks up multiple samples and optimizes parameters by looking at their slope and move forward whereas, stochastic gradient take one sample at a time. Stochastic gradient descent is considered to be faster than gradient descent.

2.7.7 Batch Normalization:

Batch normalization helps to reduce overfitting. It has relatively low regularization effects. Just like dropout, hidden layer are induced with noise using batch normalization. But batch normalization helps to reduce dropout which is a better approach because less information loss has been seen in batch normalization then dropout. But batch normalization cannot be used alone for better performance dropout and batch normalization should be used together. Batch normalization helps a network layer to learn without depending upon other layers. It reduces covariance shift by normalization of each layer's activation. More stable distributions are made for inputs thus increasing training speed. It also allows networks to use higher learning rates which improves network training speed [33].

2.7.8 ReLu Layer:

An ANN architecture necessarily has the activation function of a unit (neuron). Since the early days of ANN, researchers have been using different functions. But a good error approximation is not possible due to step function's binary nature. The sigmoid functions were utilized to overcome this challenge. They were used to provide promising results for small networks. However, it was not appropriate to scale sigmoid function on large networks [34]. As it could lead to huge numbers, the cost for computations was too high [35]. The gradient vanishing

problem was among other significant issues with sigmoid function. The prevention of learning occurs due to high value of gradient value.

Compared to previous common activation functions, the rectified linear unit function (ReLU) provided benefits in this situation. It did not suffer from the gradient vanishing problem and provided a good error approximation but it was used to cost less. The figure 2.10 displays ReLU and it is also mentioned below:

$$f(x) = \max(0, x) \tag{2.3}$$

The research conducted by Krizhevsky et al. [34] elaborated that the use of ReLU lowered the epochs's number required to converge when using Stochastic Gradient Descent by a factor of 6. There is a major drawback of ReLU's use which is the weakness when input distribution is below zero. It is due to the reason that neuron will not activate by any data point. There are practical reasons to train deep networks with the help of a GPU.

To reduce training time compared to CPU training is the main reason [35]. Though the speed depends on the network topology, the use of GPU provides 10 times faster speed [36].

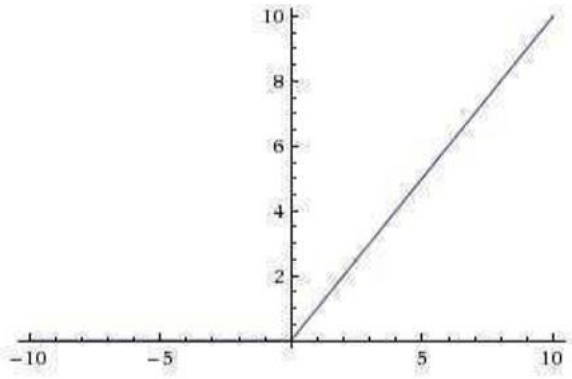

Figure 2-10: Rectified Linear Unit (ReLU) [34]

The way of processing different tasks is what differentiates GPU from CPU. A few cores are used in CPUs to execute sequential serial processing. However, GPU represents a mighty

parallel architecture. To manage several tasks at the same time, thousands of tiny cores work in harmony in this architecture.

From the above discussion, it is clear that the Deep Learning works better with GPU rather than CPU. However, most of projects still make use of CPU due to different reasons.

2.7.9 Fully Connected Layer:

After all convolutional and pooling layers high level decisions and reasoning is made with the help of fully connected layer in which each neuron is fully connected to all the neurons in the previous layer. The receptive field of a fully connected layer is the whole preceding layer. This layer is normally used to further process data for classification.

2.7.10 Loss Layer:

Loss layer is generally the final and last layer of the network which predicts the output labels and provides the error rate between predicted output labels and actual output labels and tells how much a network has been failed in training in prediction of output labels. Different loss functions are used at this layer as per the requirement. Softmax is most common loss function used to classification.

Chapter 3

Literature Review

CHAPTER3 : LITERATURE REVIEW

Emotion recognition and analysis based on the video or audio data is a research hotspot in computer vision and affective computing domain. Humans are highly social species and mostly they interact and exchange information by their face gestures. Using these gestures one can understand emotions and support all facts of life. With the dawn of new digital technologies, growth of globalization, it is needed to understand which facial expressions are related to social communication and those which cause misunderstanding should be avoided. It has been shown that understanding face expressions can alter the interpretation of what is actually spoken and these expressions control the way in which conversation should take place.

Now let us dig a little deeper in this area of research by studying and analyzing the past works, conducted by various researchers.

3.1 Classical techniques of Emotion Detection:

Wang et al. [37] investigated on kernel based methods to identify emotions from the audio-visual information. Hamming window was used, having a size of 512 points while keeping 50% overlap ratio between successive window frames. From these frames, audio features such as MFCC coefficients, pitch and power were evaluated. Face area in frames was detected in HSV color space using the Planer envelop approximation method and its parameters were tuned in order avoid false face detections. After detecting face areas, facial feature were extracted by using Gabor filters having filter bank of 5 scales and 8 orientations. To reduce the dimensionality of the extracted features, Gabor coefficients were down sampled and principal component analysis (PCA) was applied to further reduce the dimensions of the feature set and to decrease the computational complexity of the problem. Kernel Matrix Fusion (KMF) is employed to map multi-modal features into a single subspace. Kernel matrixes are developed separately for each modality, followed by the application of unsupervised Kernel principal component analysis (KPCA) or supervise Kernel Discriminant Analysis (KDA) and the transformed KPCA features are passed on to the HMM classifier. RML dataset was employed for this experiment and the proposed method outperformed other strategies including CCA, KCFA, feature level and score level fusion. Haq et al. [38] evaluated emotional states on SAVEE dataset in his audio-visual emotion recognition experiment. For that purpose, features were extracted from both modalities

separately and fusion process was tested at both feature and score level stage. Features such as MFCC, pitch, energy and duration were extracted from the audio input whereas visual features were extracted on the basis of positions of 2D marker coordinates on face areas. After extraction, distinctive features were selected using Plus l-Take Away r algorithm which is based on the Mahalanobis and Bhattacharyya distance having selection on basis of KL-divergence. After the useless features were discarded, the remaining features were dimensionally reduced by the application of PCA and LDA and were passed on to the Gaussian classifier. Feature and score level fusion strategies were put to test and the results displayed high classification accuracies, with visual recognition approach and decision level fusion delivering better accuracies than audio recognition and feature level fusion strategies respectively.

Rashid et al. [39] used spatio-temporal features obtained from the visual streams which were dimensionally reduced using Principal Component Analysis (PCA). MFCC and some prosodic features are identified as audio feature representatives. Codebook was formulated for both audio and visual features in the Euclidean space after the application of PCA and SVM was used for classification of emotional states and final class prediction was derived on basis of the prediction values coming from each classifier using Bayes Sum Rule (BSR). Visual features outperformed compared to audio but combination of both modalities generated even better results.

Moreover, Action Units (AUs), valenceand arousal space (i.e., V-A space) are proposed by Chang et al. [40] to model facial behavior.

3.2 Machine learning and CNN Based Emotion Detection:

Zhalehpour et al. [41], presented a new audio-visual dataset in his paper called BAUM-1s database. This dataset comprises of six different emotional states enacted by 31 subjects, 17 of which were female. The author used this dataset along with eNTERFACE in this audio-visual emotion recognition experiment and compared their results. Peak frame was selected in videos using maximum dissimilarity (MAXDIST) based peak frame selection which selects frames on the criteria of “maximum dissimilarity”. After that, Linear Phase Quantization (LPQ) features were extracted along with Patterns Oriented Edge Magnitudes (POEM) features as the visual set of features from these databases. Linear Phase Quantization (LPQ) is similar to Local Binary Patterns (LBP) in a way that both produce feature vectors based on local histograms. LPQ were

preferred because they offer better results compared to the well-known LBP. Furthermore, POEM features were calculated because of their robustness towards illumination changes as they neglect the pixel intensities and only consider gradient magnitudes. They also provide both local and global information compared to LBP which only provide local information. In addition to the visual features, audio features such as Mel-frequency Cepstral Coefficients (MFCC) and Relative Spectral Feature based on Perceptual Linear Prediction (RASTA-PLP) were formulated using 12 and 20 order filters respectively, with an overlap ratio of 50% and a window size of 25 msec. First and second order derivatives were calculated for the already obtained coefficients and several statistical functions were applied such that the final feature vector obtained had 675 distinct features, which were then used for classification. State of the art SVM was used as a classifier for both modalities. For video features, linear kernel was employed to neutralize the curse of dimensionality while a radial basis kernel was selected to classify audio features. The outputs from both classifiers were fused on the basis of weighted product rule where the confidence values obtained from each classifier for a video sequence are multiplied and the label of the one with the maximum product is selected.

Huibin Li *et al*. [42] proposed a novel convolutional neural network which works along deep fusion (DF-CNN). They worked on multimodalities in which they have used both 2D and 3D visual data for facial emotion recognition. This DF-CNN model is made up of three sub networks in which first is for feature extraction, second subnet is for fusion and third subnet contains softmax layer for classification. In their approach 3D face image is divided into six 2D maps which represent facial attributes. Some of those attributes are geometric shapes of the face, curvature map of the face, normal map and texture mapping of the face. These entire attribute maps are used combined as input to the DF-CNN for fusion and feature extraction. This process resulted in a 32 dimensional facial representation vector. Two ways have been opted for emotion prediction; first is use of linear SVM which uses 32-dimensional vector comprised of fused features. Second is use of softmax classification which uses six dimensional feature vector representing 6 basic emotional states. Unlike other 3D networks their network is unique in a way that it combines both feature and fusion learning in a single network. To evaluate their model they have conducted experiments on three facial datasets BU-3DFE Subset I, BU-3DFE Subset II, and Bosphorus Subset. In Bosphorus subset dataset they have only used 3D images.

Kim *et al.* [43] have performed emotional recognition on images. In which they have considered two main perspectives of an image the object and the background. According to them the meaningful information represented in image in the form of expressions is the root of emotion recognition. They have used background information in combination with object information to improve the performance of an emotion recognition system. They have extracted different features containing information of object as well as background and used those features as input to a deep neural network. The network than generates the emotion values for the image provided. They have used valence and arousal model to predict the emotions using their proposed model. They have extracted four types of features from images color features, local features, semantic and object feature and then those features have been normalized to [0, 1]. To conduct experimentation for their model they have created their own dataset and the images to make dataset were collected from Flickr. They have collected those images by keyword search and the keyword were basic emotional states i.e. angry, happy sad etc. By using these keywords initially they were able to collect more than 20,000 images from which they have manually discarded non emotional images. They have used Amazon Mechanical Turk (AMT) for assignment of values to emotions in their valence and arousal model on all the images in their dataset. The model that they have been used for training purpose contained 5 layers in total input layer, output layer and three hidden layers.

Zhang et al. [44] proposed deep belief networks (DBN) model for emotion classification. Twostage learning strategy was employed in which one stage deals with training the audio-visual network and the other one is the fusion network. The first stage involves fine tuning of the audio and visual networks through pre-trained AlexNet and C3D-Sports-1M respectively. The second stage involves training of DBN fusion network using target emotional database. 1-D audio signal is transformed into three channels of Mel-spectrogram of 64 x 64 x 3 dimensions and is passed on as input to the CNN. Size of this Mel-spectrogram can be changed conveniently in accordance with the input of the existing CNN models pre-trained on image datasets. Visual features are extracted from video segments using 3D-CNN. This process involves division of videos into segments having 16 frames, followed by detection of face in each frame of the segment by using a real-time face detector provided by Viola and Jones. After detecting face, eye distance is calculated for each frame and is normalized to a fixed value of 55 pixels. RGB image of 150 x

110 x 3 is separated from each frame base on the set value of eye distance. But the image is then resized to 227 x 227 x 3 in order to achieve fine tuning when passed on as an input to the pre-trained 3DCNN model. The DBN model used for network fusion is constructed by two RBM's stacked over each other and is trained in two steps. First step involves using greedy layer-wise training algorithm in the bottom up manner for unsupervised pre-training. After pre-training, RBMs are initialized and fine-tuned to optimize the network parameters. Audio and Video network parameters are fixed for second stage training while fusion network parameters are updated to provide accurate prediction values. After the fusion network is trained, feature representations are obtained for each audio-visual segment. Since each segment varies in length, average pooling is applied to all segment features to obtain uniform global feature representations. Different fusion strategies were studied which included feature level, score level and decision level fusion and their impact on classification rate was compared to the proposed DBN network based fusion. RML, eNTERFACE and BAUM-Is emotional datasets were selected for this experiment and linear SVM was used as the classifier. The results obtained proved highly in favor of DBN fusion network compared to the other fusion based strategies.

S.Zhang *et al.* [45] contributed towards audio emotion recognition. They proposed a deep convolutional neural network which uses discriminant temporal pyramid matching for classification of emotions in audio. In their research they have calculated MFCC features from audio files and from those feature they have extracted three channels static, delta plus delta-delta just like RGB channels of an image to make it suitable input for their proposed DCNN. These features are then divided into a number of overlapping slices. These segments are then fed into a DCNN to learn segment level features. Authors used AlexNet model which has been pre-trained on imageNet dataset for training and learning of segment level features. These segments are divided by utterances. Furthermore, DTPM has been used to combine the utterance level features and segment level features which are then classified by SVM (Linear Support Vector Machines). Four public dataset have been used for experimentation purpose namely BAUM-1s, EMO-DB, RML and eNTERFACE-05. Their proposed methodology indicated auspicious results on all four datasets.

J. Zhao *et al.* [46] proposed a deep CNN which in which they have merged two convolutional networks. Their model was comprised two branches; first branch contained 1D-CNN which has been used to learn features from raw audio clips and the second branch 2D-CNN has been used to learn MFCC features. Features extracted from both branches are the merged together for further processing followed by a fully connected layer with softmax classification. 1D-CNN contained six convolutional layers (1D), six max pooling layers (1D) and two dense layers also called fully connected. 2D-CNN contained four convolutional layers (2D), two max-pooling layers (2D) and two dense/ fully connected layers. After extracting features fully connected layers from both 1D-CNN and 2D-CNN have been deleted and the networks are then merged together. According to the author the basic purpose of developing a merged architecture was to learn various features from multidimensional data using different dimensional networks. For experimentation purpose EMO-DB and IMOCAP datasets were used. For cross validation of their proposed architecture they have used both speaker dependent and speaker independent validation techniques. Their results show that their model contributed well towards performance improvement on both datasets.

B. Yang *et al.* [47] conducted a research which focuses on weighted mixture approach. They have developed a deep neural network with weighted mixture approach and the main purpose of model was to extract affective features. They have performed a lot of preprocessing on data to extract useful features such as data augmentation, face detection and rotation etc. They have considered two channels which are being processed by their model. One channel contains grayscale images and the other contains their local binary patterns. Features from grayscale images are being extracted by fine tuning of VGG-16 network and features from LBP are extracted using a convolutional neural network. Features extracted from both channels are then fused together into weighted mixture model. Final output has been calculated via softmax classification. Three public datasets has been used for experimentation and validation of their model CK+, Oulu-CASIA and JAFFE. Their results depict that their model has performed well on all three datasets and the recognition rates achieved were among the highest ever achieved.

P. Tzirakis *et al.* [48] conducted their research on both modalities audio and visual. For audio they have divided audio signals into 6s long sequences. Then some preprocessing has been

performed to observe loudness variation in speech of different speakers. This has been done by processing the sequences in such a way that their means and unit variance is zero. They have used LSTM layers with a depth of two on top of their convolutional network to handle the temporal nature of speech. Max pooling in audio segment has been applied across time as well as across channels. For visual modality they have used a residual network with 50 layers whose input was pixel values got from cropped facial area from actor's video. Authors have performed fine tuning with a pretrained network ResNet-50. The features extracted were then trained using an RNN network having LSTM layers with a depth of two. Each layer of LSTM contained 256 cells. LSTM layers were used to handle temporal nature of data. Finally, after obtaining visual and audio features LSTM layers from both networks were removed and the feature vectors obtained from both modalities were then combined to feed into an RNN with 2 LSTM layers having 256 cells in each. For validation and experimentation they have used a spontaneous dataset named RECOLA. Results obtained from their proposed architecture shows promising improvement in recognition rates.

X. Xia *et al.* [49] performed a comparison of hand crafted features and deep learning features for emotion recognition using video data. For hand crafter features Histogram of gradient (HOG), geometric features and face shapes were calculated in their research. Whereas, for deep learning features they have fine-tuned VGG-16 model on their dataset. After fine tuning training was performed on a network having two fully connected layers. Features extracted using hand crafter methods and from deep learning are then fed into Hidden Markov Model. Probabilities of emotions obtained from HMM are then fed into Naïve Bayes classifier for classification. They have used CHEAVD 2.0 dataset for evaluation and experimentation of their proposed methodology. Recognition rates achieved by them were amongst highest ones.

H. Miao *et al.* [50] conducted their research on CHEAVD 2.0 dataset. They worked on multimodalities i.e. audio and visual data. OpenSMILE toolkit has been used in their research to extract audio features. They have extracted two types of audio features; inter speech emotion feature set 2009 (IS09) having feature vector of 384 dimensions and a feature set of large OpenSmile having a feature vector of 6552 dimensions. For traditional classification SVM, REPTree and random forest have been used. Average of scores obtained by all three classifiers is

then used for recognition. Whereas, for deep learning approaches, a Deep belief network with 5 layers with a stack of RBM has been used and its output has been used as input in multilayer perception network (MLP) for classification. The network has used back propagation for training. On the other hand from video data authors have extracted LBP –TOP features using OpenSMILE toolkit for tradition approach and for deep learning VGG CNN has been used to extract features. Hand crafted LBP features are passed to SVM classifier for classification purpose. Whereas, features extracted from CNN are passed to an RNN for classification. For multimodality authors used decision level fusion strategy. They took an average of scores obtained from both audio and visual modalities from different models. Their contributed positively towards emotion recognition challenge.

J. Zhao *et al.* [51] have multimodalities for both hand crafted features and deep learning features. Their research focuses on recognition of emotions of real time data such as TV series, talk shows and some video clips obtained from Chinese movies. They have targeted eight states of emotion; six basic emotional states along with worried and anxious being two extra emotional states. For audio, they have extracted MFCC features using OpenSMILE toolkit as traditional way and for deep learning approach they have used SoundNet deep convolutional network to extract audio features. For visual modality features have been extracted using VGG-NET, Dense-Net. Authors have also worked on contextual information in their research. In which they have extracted context features which include other objects, environment, colors and background etc. For contextual features they have used ResNet neural network. Another modality that they have worked on is textual information. To obtain text segments from video authors have utilized Jieba Chinese toolkit. These text segments are further used to obtain TD-IDF features as well as Word vector features. Word vector features are obtained from a pre trained network developed by Chinese name as Word2Vector. After extraction of all the features from different modalities LSTM network has been used training of network and to handle temporal information. For experimentation purpose they have chosen CHEAVD 2.0 dataset. Their research has contributed positively towards real time emotion recognition.

J. Y. R. Cornejo and H. Pedrini [52] contributed towards multimodal emotion recognition by means of audio and visual modalities. They have extracted MFCC features from audio modality and converted them into a segment of 64x64 by giving them a shape of image for making them suitable for convolutional neural network to process. Then they resized each portion to 227x227x3 to feed those extracted features into a pre trained audio neural network. For visual features authors have used Census transform in such a way that first thy have extracted facial area using Dlib library and then facial area from image is been cropped and the census transform has been applied to that cropped facial image. For training of visual data authors have utilized pre trained VGG-Face network. After having features from both audio and visual modality, those features have been fused together resulting in feature vector of 104448 dimensions. Authors used PCA to reduce the dimensions of resultant vector. Space created by PCA is then further processed by LDA (Linear Discriminant analysis). Finally for classification comparison they have used four algorithms namely SVM, Random Forest, Gaussian Naïve Bayes and K- Nearest Neighbor. They have used Baum-1s, RML and eNTERFACE 05 datasets for experimentation purpose. Their contribution towards emotion recognition shows promising affect in recognition rates.

I. Kansizoglou *et.al* [53] worked on dynamic learning model for online emotion recognition system. Authors have chosen audio and visual modalities for their research. Authors have proposed deep convolutional neural networks for audio and visual modalities separately and DNN for classification after fusing them together. Authors have used reinforcement learning by using long short term memory network in such a way that their system stop extracting features based on the prediction made by their network. Authors have used two datasets for experimentation of their system Baum-1s and RML. Authors have used MobileNet V2 network for visual modality by adding a layer of softmax at the end and trained the whole network on their own dataset for emotion recognition. For audio they have used VGGish pre trained network. For fusion authors have used single LSTM layer with 64 units in their network. For cross validation authors have used Leave one speaker out validation for RML and Leave one speaker Group out validation for Baum-1s. Their research has contributed positively in the field of emotion recognition and the recognition rates achieved by them were amongst the highest ever achieved. Table 3-1 summarize the latest research done related emotion recognition system.

Table 3-1: Summary of Recent Emotion Recognition Systems

Author	Dataset	Modality	Features Extracted	Classification Methodology
I. Kansizoglou *et.al* [53]	BAUM -1s, RML	Bi-modal	A: MFCC V: Distance Features, Haar Cascade base detector	A: VGGish Neural Network V: MobileNet V2 Fusion: RNN (One layer LSTM)
J. Y. R. Cornejo and H. Pedrini [52]	BAUM-1s, RML, eNTERFACE'05	Bi-modal	A: MFCC V: Census Transform, PCA, LDA	K-Nearest Neighbor, Random Forest, SVM, Gaussian Naïve Bayes
J. Zhao *et al.* [51]	CHEAVD 2.0	Multimodal	A: MFCC V: Contextual T: textual	A: SoundNet DNN V: ResNet T: Word2Vector Fusion via LSTM layer
H. Miao *et al.* [50]	CHEAVD 2.0	Bi-modal	A: inter speech emotion feature set 2009 (IS09), large OpenSmile V: LBP –TOP	SVM, VGG CNN, DBN with RBM, MLP, RNN with LSTM.
X. Xia *et al.* [49]	CHEAVD 2.0	Bi-modal	A: MFCC V: HOG, Geometric, Face Shape	VGG-16, Hidden Markov Model, Naïve Bayes
P. Tzirakis *et al.* [48]	RECOLA	Bi-modal	A: Raw Audio V: CNN base feature extraction	A: RNN with one LSTM layer V: ResNet-50, two LSTM layers
B. Yang *et al.* [47]	CK+, Oulu-CASIA, JAFFE	Uni-modal	V: LBP, data augmentation, face detection, rotation	V: VGG-16 CNN, Softmax

J. Zhao *et al.* [46]	EMO-DB, IMOCAP	Uni-modal	A: Raw audio, MFCC	A: 1D CNN, 2D CNN
S.Zhang *et al.* [45]	BAUM-1s, EMO-DB, RML, eNTERFACE-05	Uni-modal	A: MFCC	A: AlexNet DCNN, DTPM, SVM
Zhang et al. [44]	RML, eNTERFACE, BAUM-Is	Bi-modal	A: MFCC, RASTA-PLP V: CNN Based feature extraction	A: AlexNet V: C3D Sports Fusion: DBN with RBM
Kim *et al.* [43]	Created own dataset from Flickr	Uni-modal	V: color features, local features, semantic, object feature	V: CNN
Huibin Li *et al.* [42]	BU-3DFE Subset I, BU-3DFE Subset II, and Bosphorus Subset	Bi-modal	A: MFCC V: geometric map, curvature map, normal map, texture map	AV: DCNN Softmax, SVM
Zhalehpour *et al.* [41]	BAUM-1s, eNTERFACE'05	Bi-modal	A: MFCC, RASTA-PLP V: LPQ, POEM,	AV: SVM
Paleari *et.al.* [54]	eNTERFACE'05	Bi-modal	A: MFCC, LPCC Formants, Pitch, Energy V: Distance features	AV: CNN
Huang *et al.* [55]	eNTERFACE'05	Bi-modal	A: pitch, energy, speed, MFCC V: TFP, LPTP	AV: Neural Net (Genetic learning)
Fadil *et al.* [56]	RML	Bi-modal	A: MFCC, log spectrum, std, mean F0 V: FFT,PCA	AV: Deep Networks Multi-layer Perceptron
Gharavian *et al.* [57]	SAVEE	Bi-modal	A: MFCC, pitch, energy, formants V: Facial marker	A: FAMNN V: FAMNN AV: PSO optimized FAMNN

			locations	model
Noroozi *et al.* [58]	eNTERFACE'05, RML, SAVEE	Bi-modal	A: MFCC, deltas, pitch, intensity, percentile, formants, bandwidth, FBE V: Geometric features, 3D-CNN	A: SVM, RF V: SVM, RF AV: SVM, RF (with or without PCA)
Liu *et al.* [59]	RAVDESS, MNIST	Bimodal	A: MFCC, LPCC V:eigen values, CNN	AV: Deep CNN

3.3 Limitations and Gaps:

Factors that can affect the performance of the system include noise, occlusion, voice quality, illumination changes and most significantly variety of facial expressions due to different regions and disparate emotional expressions between various people. In literature the major limitations that found were related to continuous emotions and applying different fusion strategies. For instance I. Kansizoglou *et al.* [55] worked on early prediction due to which their system was not able to identify spontaneous reactions properly. Similarly S.Zhang *et al.* [45] mentioned their system's limitation in their research as not providing supervision for continuous emotions. B. Yang *et al.* [49] worked on mono channel due to which their system was unable to perform well in poor lighting condition for which multiple channels could be considered. By looking at limitations found in literature we found some research gaps which were mainly focused on supervision of continuous emotions and fusion of both modalities. In our research we tried to cover two major factors that are responsible for good recognition which are spontaneous reaction by adding time factor to our models and fusion of modalities on later stages. Most of the researchers have worked on early fusion whereas; according to recent research some researchers proposed that late fusion tend to perform better. Keeping in mind this fact from literature we have employed late fusion in our system.

Chapter 4

Proposed Methodology

CHAPTER4 : PROPOSED METHODOLOGY

This chapter presents proposed methodology of this research in detail. It comprises of techniques for extracting most distinctive features and the classification methods based on them. First we will discuss feature extraction methods followed by the classification strategies used in our study. The summary of our multi-modal approach is described in Figure 4.1. Our multi-modal emotion recognition system focuses on both audio and visual content of the video clips obtained from available emotional corpuses. Most distinctive features are extracted from both modalities separately. The following section explains this technique in detail.

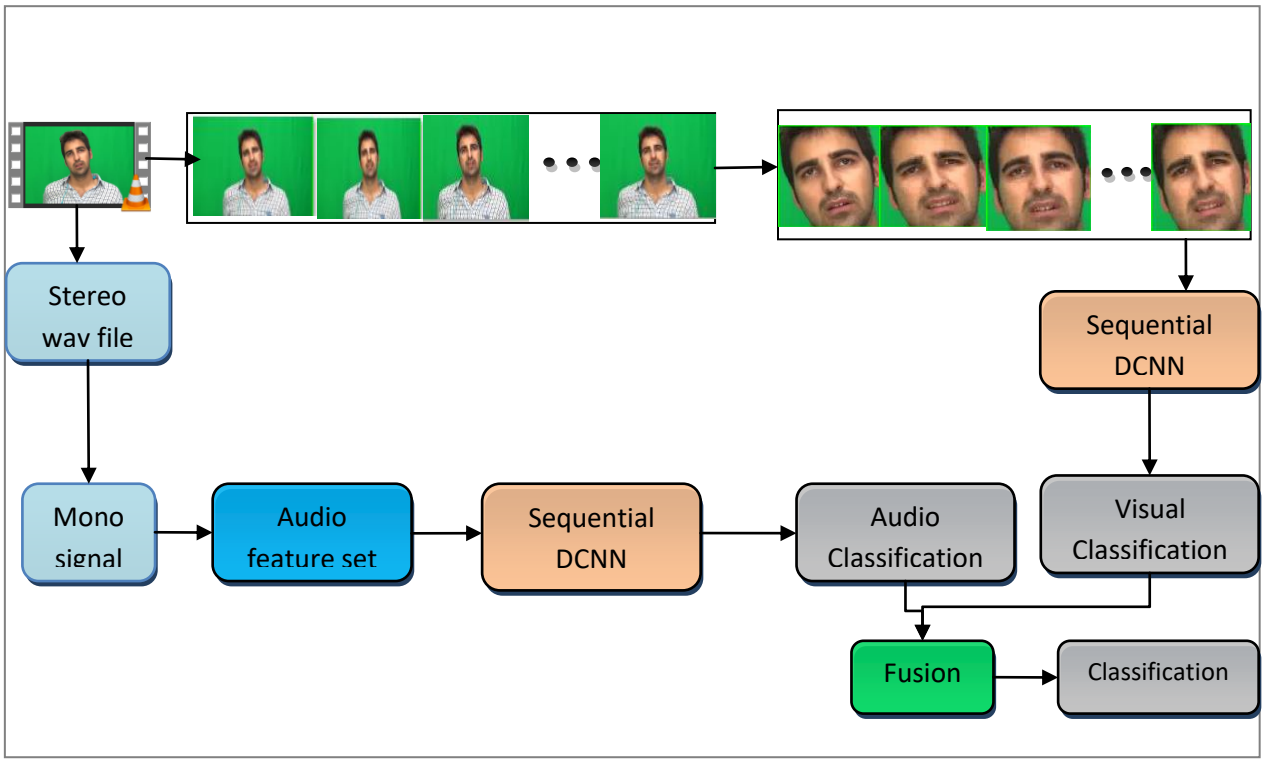

Figure 4-1: Proposed methodology for Audio-visual emotion recognition system

The first step involves some preprocessing of video data in order to make it suitable for convolutional neural network. In order to recognize emotions in a video both audio and frames were extracted separately from video for audio and visual modality to process respectively. This section describes individual details of audio and visual modality processing for emotion recognition.

4.1 Audio Modality:

The first step in recognizing emotions in audio was to extract audio wav files from video data in order to extract useful features from audio which can then be used for emotion recognition. Stereo wav files were extracted from video data having mono channel with frame rate of 44000 Hz.

4.1.1 Feature Extraction:

Emotions can be traced by feature representatives within the speech content. For example, anger can be distinguished by its high frequency, energy and rate of speech. For the same purpose, audio data is separated from the video clips and prepared for feature extraction. Many authors agree upon some of the fundamental features that provide good understanding of emotions from the speech content [60], [61]. We have utilized the commonly used Mel Frequency Cepstral Coefficients for extracting useful features.

4.1.2 Preprocessing of Audio Signal:

Before calculation of MFCC features we have done some preprocessing on the audio signal that we have obtained from video data. Following steps have been performed in preprocessing.

1. Mean signal length have been calculated by taking average of signal lengths of all the signals in audio data.

2. If signal's actual length is less the required length then signal have been padded to have the same size. Figure 4.2(b) shows sample of padded signal.

3. After that pre-emphasizing has been performed with a shift step of 1 and having a filtering coefficient of 0.98. Sample of pre-emphasizing can be seen in figure 4.2(c).

Figure 4.2 shows some preprocessing steps. Pre-emphasizing has been commonly used technique for noise reduction. It also helps to attain better quality signal by boosting the high frequency modules of a signal whereas, leaving low frequency parts of signal in their original state.

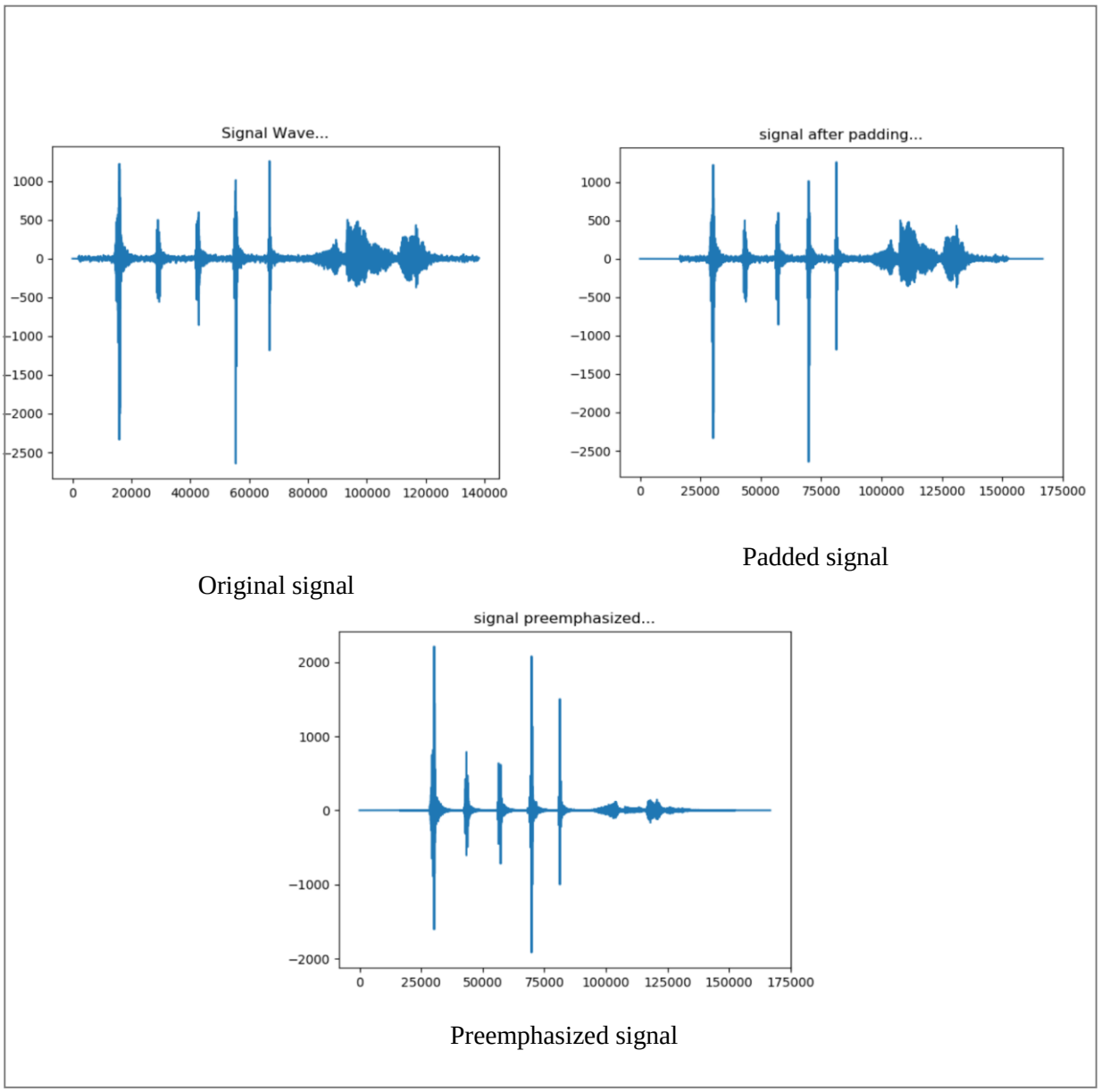

Figure 4-2: Preprocessing of audio data

4.1.3 Mel-Frequency Cepstral Coefficients:

MFCC are most commonly used features in speech recognition. They have been first introduced in 1980's by Mermelstein and Davis and have been most commonly used since then. Since they provide the closest approximation to human auditory response compared to other linearly distributed frequency bands in any ordinary spectrum, they are adopted and used in this research as well.

Steps involved in calculation of MFCCs are:

1. Frame the processed signal into short overlapping frames with window length of 25 ms and an overlap f 10ms. Figure 4.3 shows a windowed signal.
2. Calculate Fast Fourier Transform (FFT) of every frame to calculate periodogram for each frame. FFT spectrum of windowed signal can been in figure 4.4.
3. Power spectrum for each frame has been calculated with N_FFT=512.
4. Calculate Mel-filter bank energy features for all the frames with 40 filters and sum the energies.
5. Take logarithm of all the Mel-filter bank energies that have been calculated.
6. DC elimination has been performed by calculating DCT of log filter bank energies.
7. 13 coefficients were kept rest discarded. A Sample of MFCC is displayed in figure 4-5.

A toolkit named speechpy has been used to calculate MFCC coefficients. Figure 4-3 to 4-6 shows few of steps performed in calculation of MFCC. The frequency at which signal was acquired was 44000 kHz. Figure 4-6 represents MFCC spectrum with and without DCT elimination.

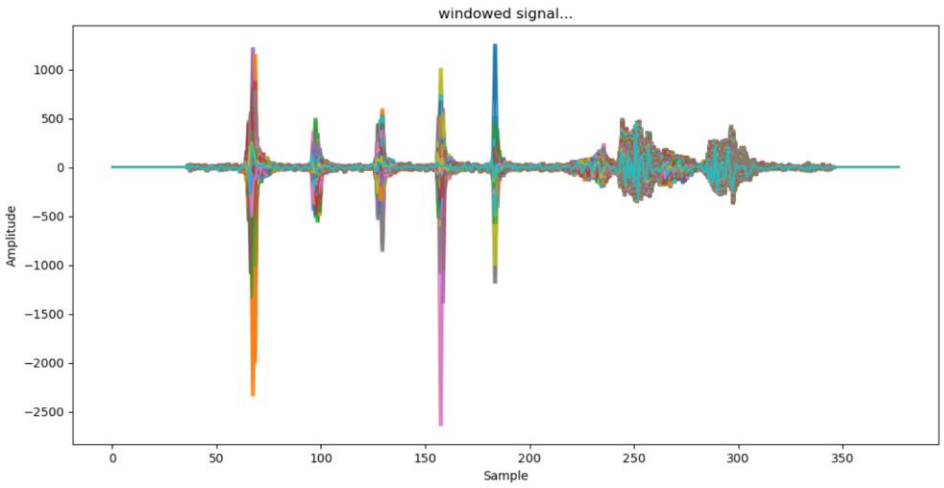

Figure 4-3: Framing of Signal

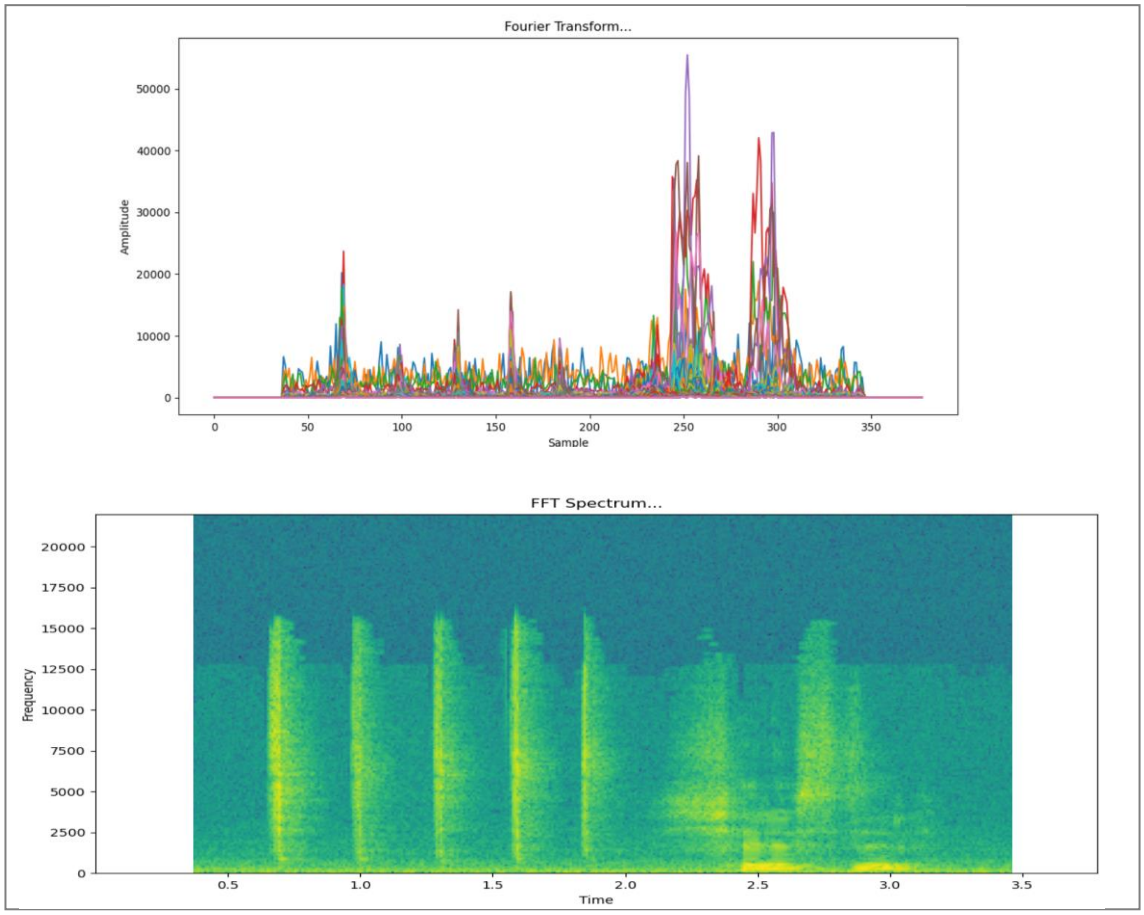

Figure 4-4: Fourier Transform and FFT spectrum of the signal

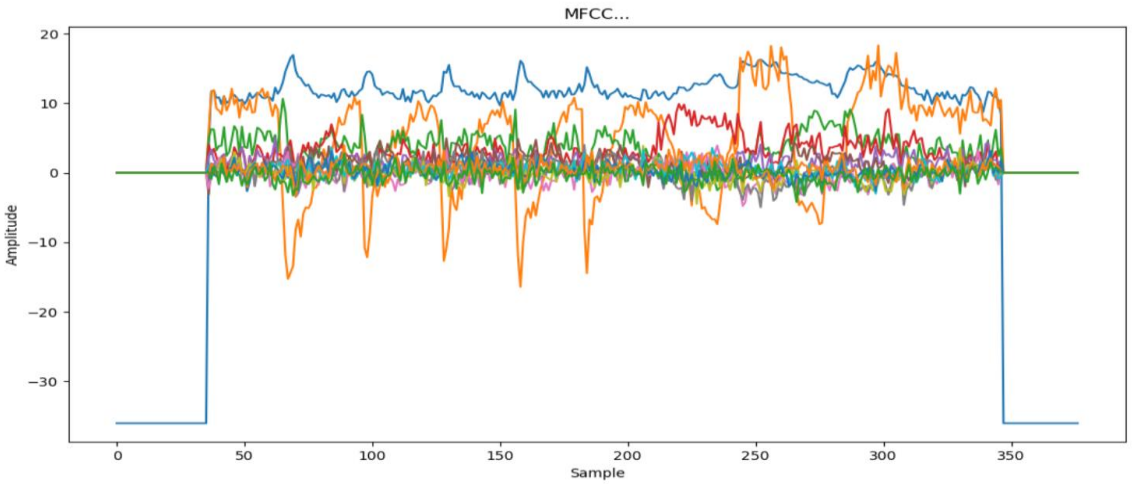

Figure 4-5: MFCC of signal

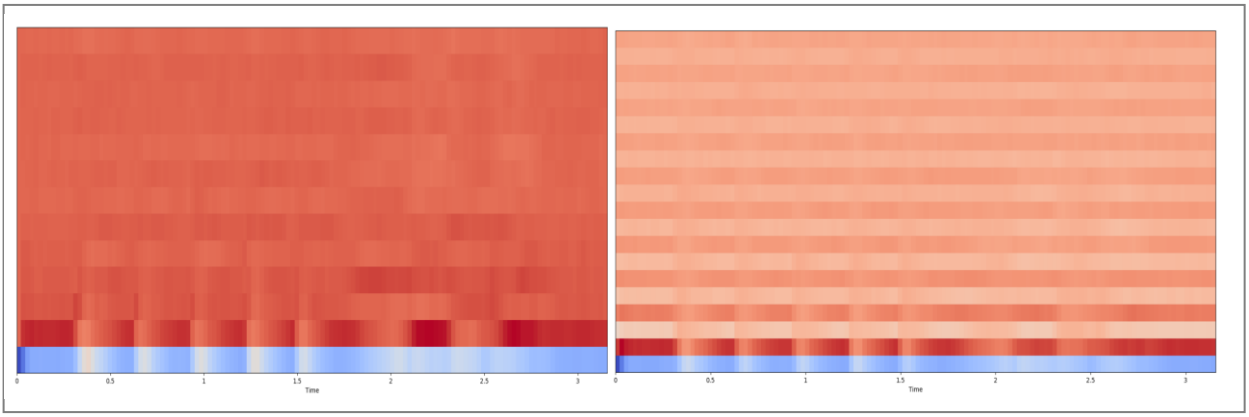

Figure 4-6: MFCC Spectrum without and with DCT

4.1.4 Mel Spectrogram:

Spectrogram is visual representation of audio which is built by stacking of FFT of a segmented audio. Audio signal is segmented into overlapping windows and for each window FFT is calculated. After calculation of spectrogram mel scale is applied to spectrogram. A human ear can hear frequency ranging between 500-1000Hz easily but it is very difficult for a human ear to listen to higher frequencies. For that purpose frequencies are converted to a scale that is audible for the listener. Mel spectrogram is combination of both spectrogram and mel scale. Mel spectrogram is basically a procedure to convert frequencies to mel scale. According to the author of [44] and [68] mel spectrogram are considered to be more efficient that other audio feature such as MFCC so we have also calculated mel spectrogram of audio signal with 64 filter banks with a values ranging from 20Hz to 8000Hz. After that log mel spectrogram is being split in to chunks with a window of 64 frames with 30 hops. Size of segmented window was 25 ms with an overlap of 10 ms. As a result we obtained 64x64x3 image shape data. Which was then resized to 227x227x3 to make it suitable for pre-trained AlexNet network. Figure 4.7 shows a sample of mel spectrogram for different emotions. Number of FFT points and hop length was calculated by using following formula. Here sr is sampling rate of audio signal which was 44000 kHz.

$$N\_FFT= (0.025)* sr \quad (4.1)$$

$$Hop\ length= (0.01)* sr \quad (4.2)$$

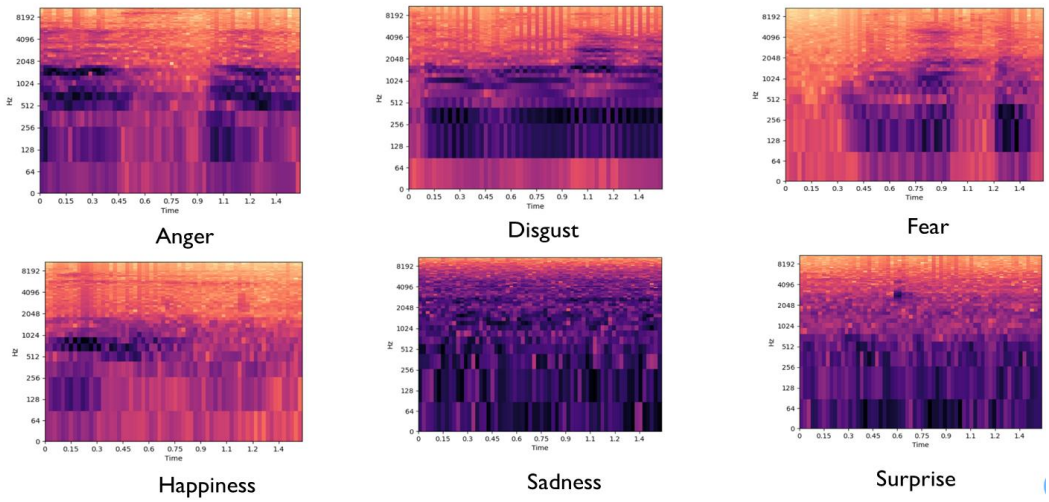

Figure 4-7: Mel-spectrogram of a wav files for different emotions

4.2 Visual Modality:

In order to perform emotion recognition in visual modality some preprocessing on visual data has also been performed. After extraction of frames from each video haar cascade based on idea of Voila Jones algorithm was used to detect facial area. Haar Cascade is an algorithm which is commonly used for object detection in visual data. OpenCV toolkit was used for face detection because it provides pre-trained haar cascade classifier for detection of face. After detection of faces images were cropped to facial area and resized to 224x224x3 to make them suitable for deep CNN. Figure 4.8 shows some samples from visual preprocessing.

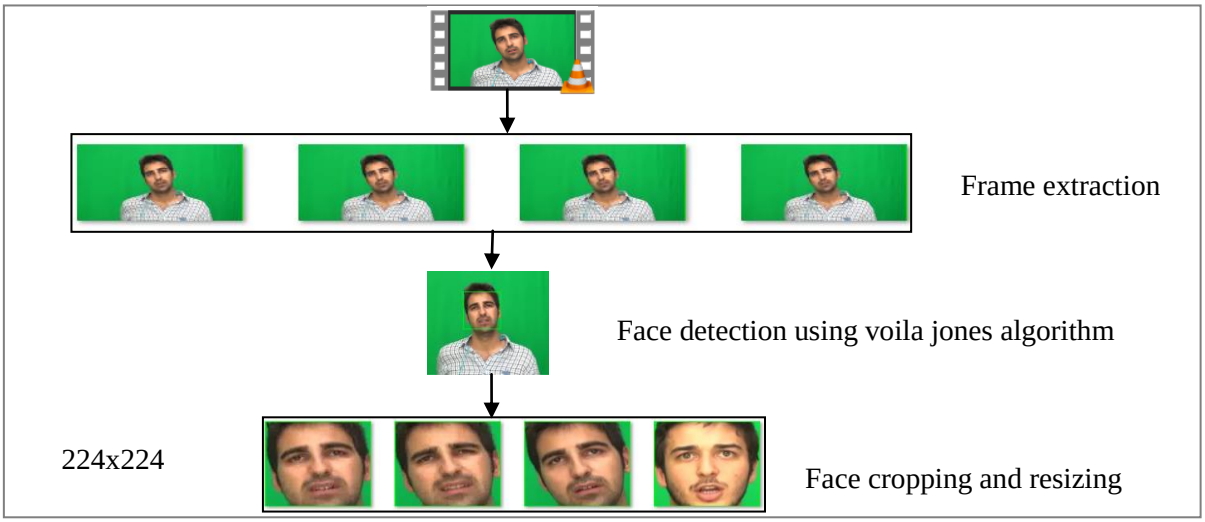

Figure 4-8: Visual data preprocessing

4.3 Data Normalization:

Data normalization on visual and audio data was performed as well in terms of feature scaling. This normalization is generally performed to scale data values between 0 and 1. Feature scaling can be achieved by rescaling of data and is implemented as follows:

$$x_{normalized} = (x - x_{minimum}) / (x_{maximum} - x_{minimum}) \quad (4.3)$$

Where $x_{normalized}$ is the normalized value and x is the original dataset.

Using this normalization training data for pre-trained ALexNet was also normalized. At the end a feature vector of 4096-D was extracted at FC-7

4.4 Deep Convolutional Neural Network for Audio Modality:

Sequential deep convolutional neural network ALexNet was used for audio classification on audio dataset which contains five convolutional layers having 96 filters of 11x11 in the first layer followed by max pooling layer of 2x2 with relu acticvation, 256 11x11 in the second layer followed by max pooling layer of 2x2 with relu activation, 384 filters of 3x3in the third layer with relu activation, fourth convolutional layer having same parameters and 256 3x3 in the fifth layer followed by max pooling of 2x2. After max pooling features were flatten to pass them to fully connected layers with 4096 neurons in fully connected layer 6 and 7 with relu activation. Dropout of 0.5 was used after both layers to avoid overfitting. After finetuning of features top head was built by adding average pooling layer followed by fully connected layer with 1000

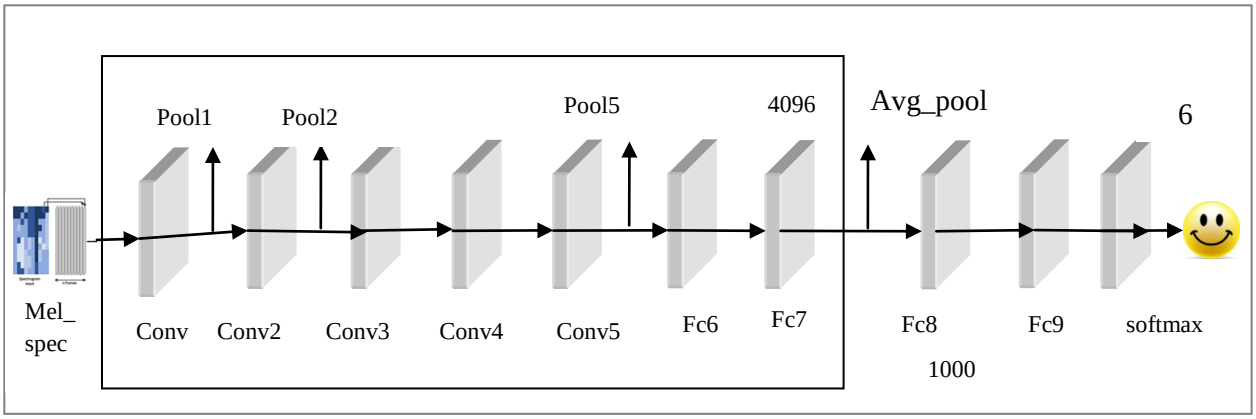

Figure 4-9: Proposed Methodology for Audio Classification

neurons with relu activaion. Dropout was also used after this layer. Finally the last fully connected layer was added which led to classification of 6 emotions using softmax classifier.

Figure 4.9 represents proposed model for audio classification. Several numbers of features are chosen for good quality of emotion detection. The Total Parameters in proposed network are 32,387,870.

Total parameters: 32,387,870 Trainable parameters: 32,387,870 Non-trainable parameters: 0

4.4.1 Pre-trained ALexNet Network:

Pre-trained AlexNet was chosen for feature extraction from mel spectrogram. We selected pre-trained parameters of ImageNet database. The architecture of AlexNet is explained in detail and shown in figure 4.10.

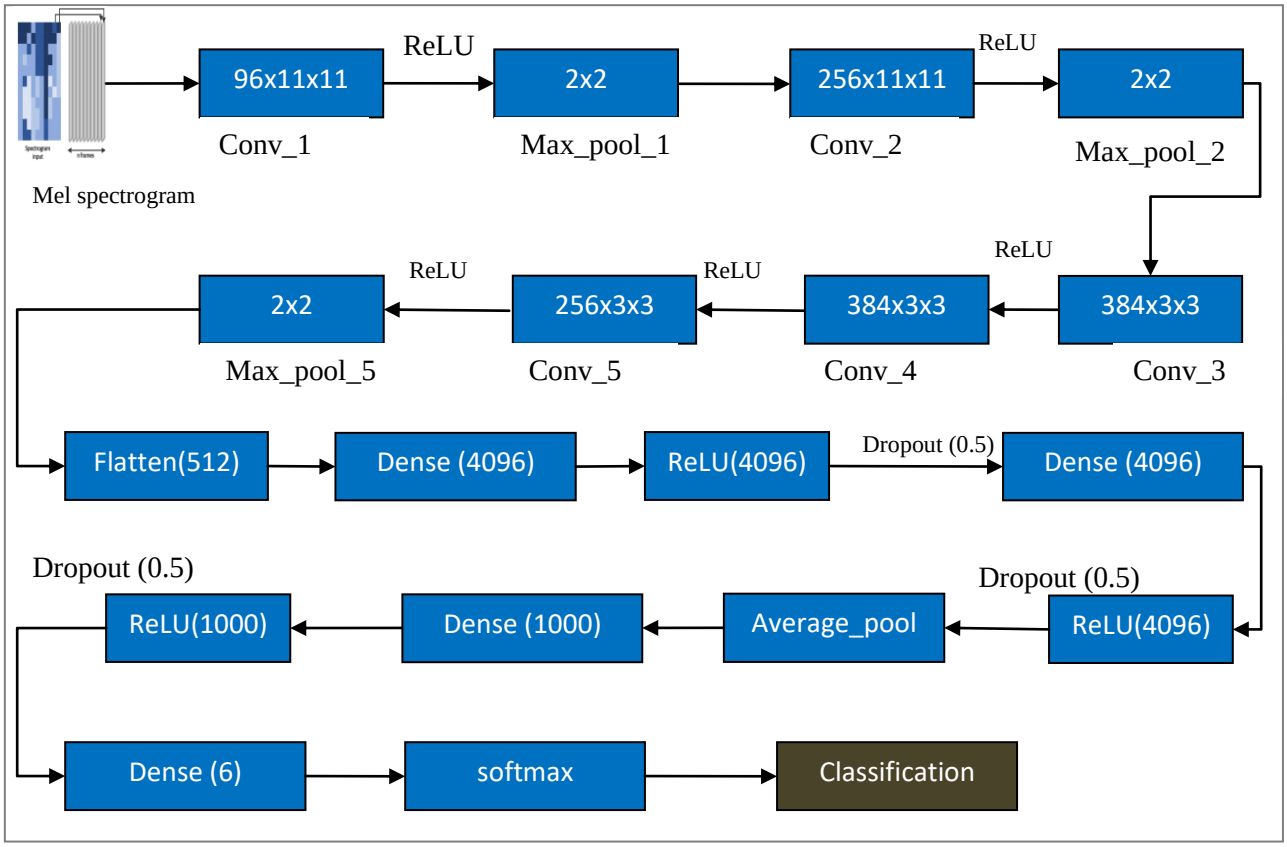

Figure 4-10: AlexNet architecture for Audio Classification

4.5 Deep Convolutional Neural Network for Visual Modality:

A deep convolutional neural network was used for emotion recognition in visual modality in collaboration with RNN having one LSTM layer. The network that has been proposed is highly inspired by VGG-16 architecture which was winner of ILSVR 2013 challenge. The network is comprised of total 5 blocks. The first block contains two convolutional layers; first with a kernel size 7x7 having 32 filters followed by batch normalization and relu activation layer, second with a kernel size of 3x3 having 32 filters followed by batch normalization and relu activation and after them a 2x2 max pooling layer has been added. The remaining four blocks follows the same patterns except their first convolutional layer contain a filter of 3x3 as well as second convolutional layer. These blocks are called with 64, 128, 256, 512 filters gradually.

After the successful extraction of useful features data has been flatten and fed into a LSTM layer with 256 units. Dropout of 0.5 has been added to minimize overfitting.

This LSTM layer was followed by fully connected layer with led the network to emotion classification using softmax classifier. The lstm layer has been used to handle the temporal nature of data. Several numbers of features were being chosen for good emotion detection. Time distributed wrapper has been used to handle the temporal slices of input data.

Time distributed layer takes several input at a time and applied similar layers to all the inputs slices and generates only one output instead of generating different output and feature vectors for every single input. LSTMs are considered to be good for handling time series data that's why many to many LSTM has been used in our model for labeling each frame of the video w.r.t emotion. Figure 4.11 represent a basic flow of many to many LSTM.

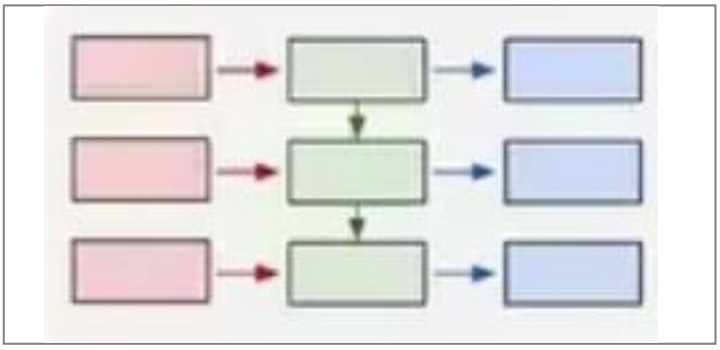

Figure 4-11: Many to Many LSTM Depiction [48]

Every rectangle in the above picture represents a vector. Red box represents input, whereas green box depicts LSTM states and blue boxes represents the output. Synced input and output sequences have been used because we want to label each frame of the visual data (Video). Proposed model architecture for visual modality can be seen in figure 4.12. Detailed summary of model has been given in table 4.1 with input and output shapes and number of parameters extracted at each layer.

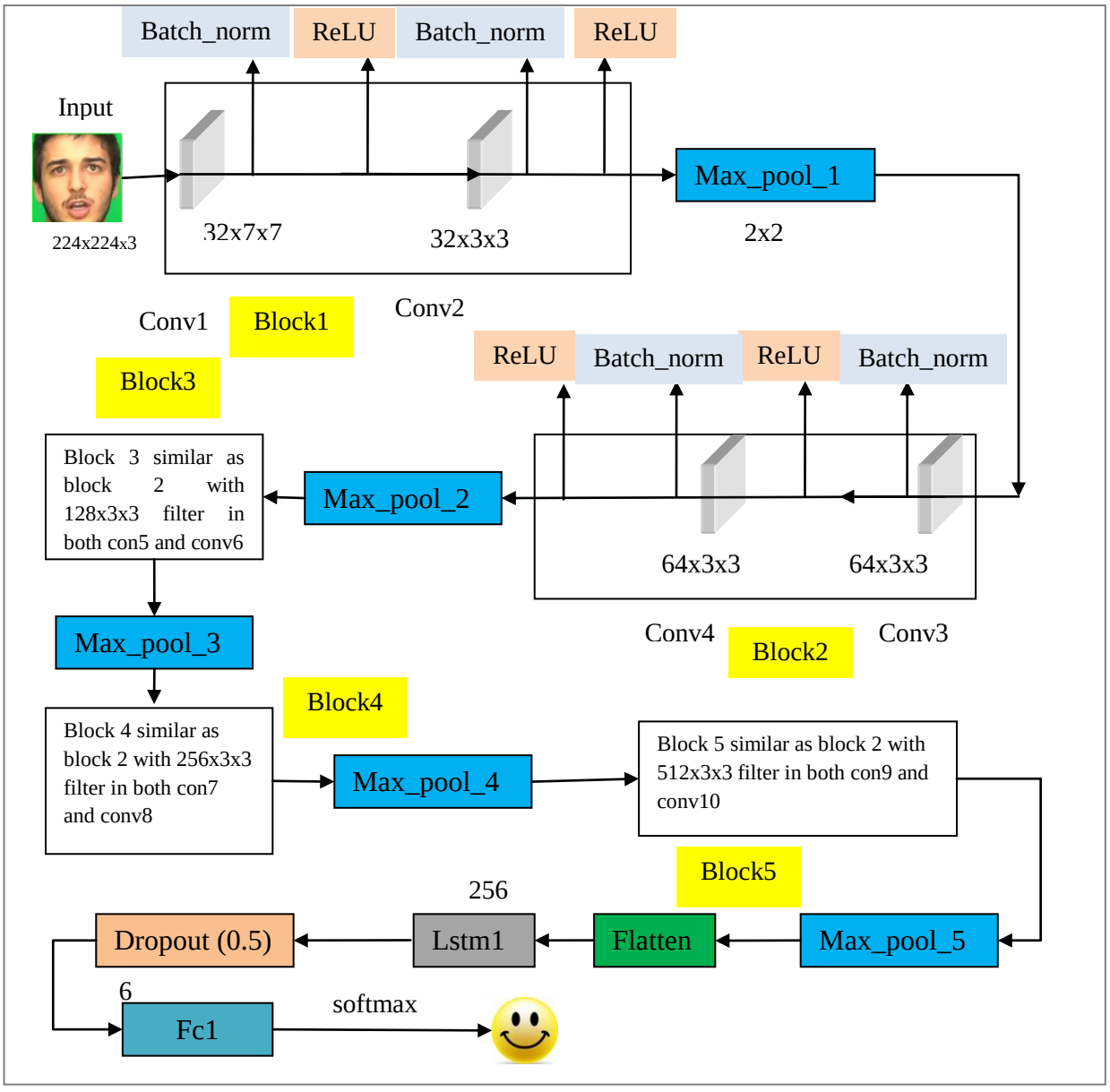

Figure 4-12: Proposed Model Architecture for Visual Modality

L2 regularization has been used in the network to reduce overfitting due to its invariant nature towards rotation and scaling. The proposed network extracted many useful features for better emotion detection. Summary of network can be seen in table 4-1.

Table 4-1: Complete Specification of each layer of Proposed Model

LAYER (TYPE)	FILTER SHAPE	OUTPUT SHAPE	PARAM #
Input_1(input layer)		224x224x3	0
Conv2d_1(Convolution)	32x7x7	32x112x112	4736
Batch_Normalization		32x112x112	128
Conv2d_2(Convolution)	32x3x3	32x110x110	9248
Batch_Normalization		32x110x110	128
Max_pool_1	2x2	32x55x55	0
Conv2d_3(Convolution)	64x3x3	64x55x55	18496
Batch_Normalization		64x55x55	256
Conv2d_4(Convolution)	64x3x3	64x55x55	36928
Batch_Normalization		64x55x55	256
Max_pool_2	2x2	64x27x27	0
Conv2d_5(Convolution)	128x3x3	128x27x27	73856
Batch_Normalization		128x27x27	512
Conv2d_6(Convolution)	128x3x3	128x27x27	147584
Batch_Normalization		128x27x27	512
Max_pool_3	2x2	128x13x13	0
Conv2d_7(Convolution)	256x3x3	256x13x13	295168
Batch_Normalization		256x13x13	1024
Conv2d_8(Convolution)	256x3x3	256x13x13	590080
Batch_Normalization		256x13x13	1024
Max_pool_4	2x2	256x6x6	0
Conv2d_9(Convolution)	512x3x3	512x6x6	1180160

Batch_Normalization		512x6x6	2048
Conv2d_10(Convolution)	512x3x3	512x6x6	2359808
Batch_Normalization		512x6x6	2048
Max_pool_4	2x2	512x3x3	0
Flatten()		4608	0
Lstm_1(LSTM)		256	4981760
Dense_1(Dense)		6	1542
Softmax		6	0

Several numbers of features are chosen for good quality of emotion detection. The Total Parameters in proposed network are 9,707,302. Specification of complete architecture is shown in Table 4-1.

Total parameters: 16,331,239

Trainable parameters: 16,331,239

Non-trainable parameters: 0

4.5.1 Long Short-Term Memory (LSTM):

Long Short Term Memory networks – usually just called "LSTMs" are a special kind of RNN, capable of learning long-term dependencies. LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn! The first step in our LSTM is to decide what information we're going to throw away from the cell state. This decision is made by a sigmoid layer called the "forget gate layer." It looks at $_{ht-1}$ and x_t, and outputs a number between 0 and 1 for each number in the cell state $_{Ct-1}$. A 1 represents "completely keep this" while a 0 represents "completely get rid of this." The next step is to decide what new information we're going to store in the cell state. This has two parts. First, a sigmoid layer called the "input gate layer" decides which values we'll update. Next, a tanh layer creates a vector of new candidate values, $C_{\sim t}$, that could be added to the state. In the next step, we'll combine these two to create an

update to the state. It's now time to update the old cell state, C_{t-1}, into the new cell state C_t. Previous steps already decided what to do, and we just need to actually do it. We multiply the old state by f_t, forgetting the things we decided to forget earlier. Then we add $i_t * C_{\sim t}$. This is the new candidate values, scaled by how much we decided to update each state value. Finally, we need to decide what we're going to output. This output will be based on our cell state, but will be a filtered version. First, we run a sigmoid layer which decides what parts of the cell state we're going to output. Then, we put the cell state through tanh (to push the values to be between −1 and 1) and multiply it by the output of the sigmoid gate, so that we only output the parts we decided to.

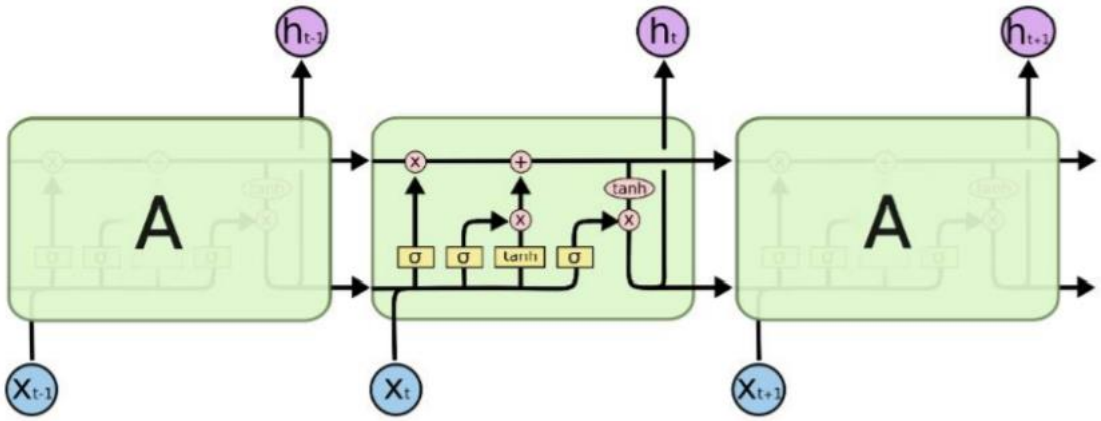

Figure 4-13: LSTM Module [51]

4.6 Neural Network Parameters:

The proposed architecture is implemented using Keras library in python. These libraries provide numerous methods and models for implementing CNN in Tensor-flow as backend. CNN network in Keras can be implemented in both CPU and GPU. We have Google Colab GPU having 1xTesla K80, compute 3.7, having 2496 CUDA cores, 12GB GDDR5 VRAM for training and testing purpose.

In network those parameters which can be tuned are tuned for the best possible results. ReLU and Soft-max are used as neuronal activation parameters, stochastic gradient descent (SGD) as optimization parameter and Glorot uniform as initialization parameter for different layers of network. Batch size was fixed to 32 and number of epoch was 10. Bias value was fixed to be

zero. Stride of 2 was used for first convolutional layer. Stride of 2 was also used for each max pooling layer.

4.6.1 Neuronal Activation:

The purpose of the neuronal activation function is to control the output of the neurons in neural network. There are many types of neuronal function e.g maxout, RELU, tangent etc. The one which we have used in our proposed network is RELU which is a nonlinear function. The output of RELU function is only two values, zero and a positive value. This function finds the max value between zero and the given output.

$$F(x) = \text{Max}\ (0\ ,\ x). \tag{4.4}$$

4.6.2 Regularizer:

In neural network when system neuron learns, they try to tune their weights according to input data. Neighbor neurons start to depend on each other for specific data and if this goes on system become over fitted for specific data which results in a fragile model. To address this problem Regularizers are used. Their purpose is to reduce the overfitting in model. Dropout and L2 regularizer has been used in our network. Dropout Regularizer selects neurons randomly and drops them. This way dependency of neurons on each other is minimized. Remaining neurons will have to update their weights independently. In our proposed neural network high dropout value of 0.5 is used throughout the network to reduce the chance of overfitting. L2 regularization has also been used in the network to reduce overfitting due to its invariant nature towards rotation and scaling. L2 regularizer takes the squares of weights and calculates their sum. L2 forces the weights to be so small and more robust.

4.6.3 Optimizer:

Purpose of the optimizer is to distribute weights throughout network by using the loss function at the output of network. Stochastic gradient descent (SGD) [35] is used as optimizer and Categorical Cross entropy as loss function in our network. Categorical cross entropy is used because we are dealing with multi class classification problem.

4.7 Post processing and Fusion:

Once we had classification results from both modalities i.e. audio and visual data. We moved towards fusion of both modalities. Two types of fusions were performed in order to get the results of combined modalities.

4.7.1 Decision level Fusion:

Decision level fusion is performed in such a way that it takes recognition accuracies from each modality and combines them according to an algebraic principle. The authors of [70] have defined five decision level fusion rules which are Min, Max, Sum, Median and Product rule.

- **Max Rule** is defined by following expression:

$$d_j(X) = \max_{i=1}^{l} c_{i,j}(X), \quad j = 1, 2, \ldots, m. \tag{4.5}$$

  Where i denotes the classifier and j denotes the recognition rate of emotion classified by i^{th} classifier. Max rule picks up the maximum value classified as the fusion output.

- **Min Rule** on the contrary selects minimum value classified as fusion output. Min rule through an exprression can be defined as:

$$d_j(X) = \min_{i=1}^{l} c_{i,j}(X), \quad j = 1, 2, \ldots, m. \tag{4.6}$$

- **Sum rule** can also be defined as the mean. It simple takes the mean of recognition rate classified by different classifiers. It can be defined as:

$$d_j(X) = \sum_{i=1}^{l} c_{i,j}(X), \quad j = 1, 2, \ldots, m. \tag{4.7}$$

- **Median rule** calculates the median of recognition score of every classifier. If the classifiers are even in number then resultant is mean of the two medians.

It can be expressed as:

$$d_j(X) = \underset{i=1}{\overset{l}{\text{median}}}\, c_{i,j}(X), \quad j = 1, 2, \ldots, m. \tag{4.8}$$

- **Product rule** is simply the product of recognition score obtained by classifier. It can be summarized as:

$$d_j(X) = \prod_{i=1}^{l} c_{i,j}(X), \quad j = 1, 2, \ldots, m. \tag{4.9}$$

All types of decision level rules are applied in our research. Experimental results of fusion strategies are given in chapter 5. Highest value obtained from combined feature vector by particular was considered to be emotion label.

4.7.2 Score Level Fusion:

Score level fusion is a modified version of decision level fusion which has been used recently by researchers [71][72]. As per the author of [71] sum of classification score is calculated with equally assigned weights of each modality. As a result, the emotion class which leads to highest value is being predicted as final output. Score level fusion is performed through summation of individual scores obtained from each classifier which depicts the probability of occurrence of emotion for a sample. Whereas, decision level fusion is performed by integrating multiple labels predicted by classifiers [44]. We have employed score level fusion as per the scheme in [44][71].

$$Score^{fusion} = 0.5Score^{audio} + 0.5Score^{visual} \tag{4.10}$$

4.7.3 Score level Fusion Classification:

Confidence score obtained from both modalities was combined to make a single feature vector. That feature vector was used as input to multiple classifiers i.e. SVM, Random Forest Classifier and K-Nearest Neighbor.

4.7.3.1 Support Vector Machines:

We employed Support Vector Machines (SVM) in our research for classification of audio and video features. Multi-class SVM was used which is a combination of multiple binary SVM [40].Hence M-SVM is adopted in our experiment which requires classification among six or

more emotional states. On providing labeled training data, SVM algorithm generates an optimal hyper plane which separates and categorizes different observations. SVM is known for its capability to solve high dimensional problems with high efficiency and better recognition accuracy [25]. As we have extracted features from audio and videos, our problem is dimensionally complex and therefore SVM is one of the optimal solutions.

4.7.3.2 Random Forest:

We also experimented with Random Forest (RF) classifiers in our recognition model. RF classifiers are known for their simplicity and accuracy. It is an ensemble of tree-type classifiers $\{h(x, \theta_k), k=1,..,\}$where x is the input vector and $\{ \theta_k \}$are independent and identically distributed random vectors [57]. In a classification problem, single vote is cast by each tree to determine the popular class at input and the output is determined by the majority voting principle as shown in Figure 4.5. The algorithm is not sensitive to the number of variables although they are a user-defined parameter. However, these values are normally set to the square root of the number of inputs. This not only reduces the computational complexity of the algorithm by a fair margin but also decreases the correlation between the trees. Moreover, trees of the forest are not pruned which further reduces the processing load.

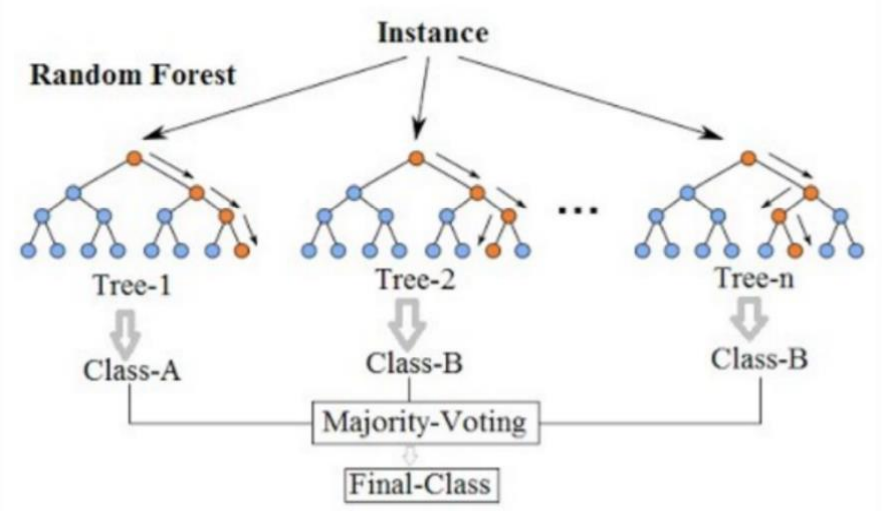

Figure 4-14: Random Forest classifier [57]

RF algorithm does not over-fit and requires no guidance whatsoever. It can also detect outliers which may prove effective if the data has mislabeled information. Owing to their simplicity, speed and powerful algorithm, they have been used in solving our emotion recognition problem.

4.7.3.3 K-Nearest Neighbor:

KNN is simplest algorithm which keeps all the available data points and classify new data points based upon distance measure. KNN has been used since 1970 for pattern recognition statistical approximation. KNN can be used for classification and regression as well. Particularly considering classification problem, KNN assumes that similar data points exist near to each other. KNN algorithm involves following steps:

1. Find out the value of K. it could be any integer. Using general rule of thumb k can be calculated according to this equation.

$$\mathbf{k} = \text{sqrt}(N)/2 \tag{4.11}$$

2. For each point in the dataset:
    a. Distance between each test data point and every row training data point is calculated using any of distance function i.e. Euclidean or Manhattan etc. The most common distance function being used is Euclidean distance.
    b. Sort the data points based on resultant calculated in step a.
    c. Top rows will be selected from sorted data.
    d. Most frequent class label will be assigned to the test data point.

Chapter 5

Experimental Results

CHAPTER5 : EXPERIMENTAL RESULTS

In this chapter, we display the outcome of our experiments conducted in our multi-modal emotion recognition system. We employed two datasets, namely, BAUM-1s and RML. Separate results for each modality have been presented individually.

These two datasets are comprised of 1264 video clips. BAUM-1s database contains 544 videos recorded by 31 different subjects in Turkish language. Whereas, RML database contains 720 videos recorded by 8 different subjects in six different languages.

5.1 Databases Used:

Several recent audio-visual emotion recognition methods have employed acted and spontaneous databases. We have further tried to contribute in that aspect by using RML as acted dataset and BAUM-1s as spontaneous dataset in our research.

5.1.1 BAUM-1s A Spontaneous Database:

BAUM-1s is a spontaneous dataset contains 1184 videos depicting 13 different emotional states. Among these 13 emotional states only six basic emotional states namely Anger, Disgust, Fear, Happiness, Sadness, and Surprise were selected for our experimentation purpose which reduced dataset size to 544 videos. The dataset has been recorded by 31 subjects, 13 of them are female and 17 are male subjects. Dataset has been recorded in Turkish language and all the actors are native speakers of Turkey. Speakers' age is ranging from 19 to 65 years. Average length of video clips is 1.82 seconds. Videos are available in two formats MP4 and AVI. This database is considered to be very challenging because of its recording circumstances and limitations. According to the creators [63] of dataset, the dataset is challenging because speakers were known of cameras so some of them might have suppressed their emotions and some have exaggerated them. Five annotators were selected to label the data. Annotators were given the video clips to watch and every annotator was requested to give score from 0-5 based upon intensity of emotion. Then as per majority labels were assigned to the video clips. Kappa value for BAUM-1s dataset is 0.54. Kappa coefficient is used to calculate the agreement among two annotators who are required to classify N number of videos for C number of categories.

The formula to calculate kappa value is:

$$\kappa = \frac{P_o - P_c}{1 - P_c}, \tag{5.1}$$

P_0 represents observed agreement and P_c depicts the hypothetical probablity for agreement. Kappa value lies between -1 to 1 [63]. Where 1 is the perfect value as it refers to agreement between all the annotators and 0 is the worst values as it refers to disagreement among all the annotators except by chance. Here are some samples from BAUM-1s dataset in figure 5-1.

Figure 5-1: Samples from BAUM-1s Dataset [44]

5.1.2 RML Acted Database:

RML dataset was recruited in our emotion recognition system. This dataset was created at Ryerson Multimedia Lab and contains 720 video clips, each having an average duration of 3 to 6 seconds and include six basic emotional sates (anger, disgust, fear, happiness, sadness and surprise) [16]. Recording were made with a digital camera in a quiet environment with a clear background and bright surroundings. Moreover, emotions are expressed in a variety of languages by people belonging to different ethnicity and background. These languages include English, Mandarin, Italian, Punjabi, Urdu and also contains some of the basic intonations regarding English and Chinese. Videos were recorded at a speed of 30 FPS and frequency of 22050 Hz, by using a 16- bit single channel digitization technique. Figure 5-2 shows samples of RML dataset.

Figure 5-2: Samples from RML Dataset [16]

5.2 Evaluation for Both Modalities:

After the detailed description of the databases in the previous section, we now move ahead and explain how the experiment was performed and the results were evaluated.

5.2.1 Cross Validation:

Leave One Speaker Out (LOSO) cross validation was performed in our experiments for RML database according to which dataset was split in such a way that one speaker was kept out for testing and others were used for training. For instance, as RML contains 8 speakers, 1 speaker was kept out for testing and system was trained on remaining 7 and same procedure was being followed for all the speakers one by one separately and results obtained from each speakers were averaged to find out the final outcome. Similarly Leave One Speaker Group Out (LOSGO) cross validation was used for BAUM-1s dataset with five speakers in one set similar as [41].

5.3 Audio Modality Results:

For emotion recognition in audio MFCC features were obtained from audio data that was obtained from original videos. 13 features were kept for each frame and remaining discarded. Those MFCC were then passed to a CNN for audio classification. Along with handcrafted features, mel spectrograms were also being used as input to pre-trained ALexNet for audio feature extraction. LOSO and LOSGO cross validation was performed for RML and BAUM-1s respectively. Table 5-1 display confusion matrix of speaker independent audio modality with MFCC features on BAUM-1s dataset. Average recognition rate with MFCC is 43.6%. Values in diagonal shows correctly classified emotion samples. Table 5-2 represents confusion matrix for Pre-trained AlexNet architecture with mel-spectrogram as input for BAUM-1s. Recognition rate achieved from pre-trained AlexNet is 46.23% on BAUM-1s dataset. It can be seen that most misclassified emotion in both models is fear whereas; happiness and sadness were most classified emotions. Table 5-3 represents confusion matrix of speaker independent audio modality with MFCC features on RML dataset. Average Recognition rate achieved was 65.36%. Table 5-4 represents confusion matrix for Pre-trained AlexNet RML. Recognition rate achieved from pre-trained AlexNet is 68.43% on RML dataset. Pre-trained AlexNet model give better recognition rates. Table 5-5 represent comparison of some recent studies with proposed methodology results on audio modality for BAUM-1s and RML.

Table 5-1: Confusion matrix for best results on BAUM-1s for audio modality with MFCC

BAUM-1s	Anger	Disgust	Fear	Happiness	Sadness	Surprise	RR(%)
Anger	9	9	4	2	11	4	23.0
Disgust	11	32	8	12	17	6	37.2
Fear	2	8	10	0	9	9	26.3
Happiness	7	21	4	120	17	10	67.0
Sadness	10	20	4	5	80	20	57.5
Surprise	2	4	13	0	2	22	51.1
						Average Recognition Rate	**43.6 %**

Table 5-2: Confusion matrix for best results on BAUM-1s for audio modality for pre-trained AlexNet model

BAUM-1s	Anger	Disgust	Fear	Happiness	Sadness	Surprise	RR(%)
Anger	11	7	4	2	11	4	28.2
Disgust	11	33	8	12	16	6	38.3
Fear	2	8	11	1	8	8	28.9
Happiness	7	21	4	122	15	10	68.2
Sadness	9	17	4	5	84	20	60.4
Surprise	2	4	12	0	2	23	53.4
						Average Recognition Rate	**46.23 %**

Table 5-3: Confusion matrix for best results on RML for audio modality with MFCC

RML	Anger	Disgust	Fear	Happiness	Sadness	Surprise	RR (%)
Anger	85	7	4	2	4	18	70.8
Disgust	2	80	13	17	7	1	66.6
Fear	3	13	65	16	16	7	54.1
Happiness	5	20	5	80	9	1	66.6
Sadness	3	9	13	6	88	1	73.3
Surprise	17	3	21	4	2	73	60.8
						Average Recognition Rate	**65.36 %**

Table 5-4: Confusion matrix for best results on RML for audio modality using pre-trained AlexNet model

RML	Anger	Disgust	Fear	Happiness	Sadness	Surprise	RR (%)
Anger	90	6	4	2	3	15	75.0
Disgust	2	83	12	16	6	1	69.1
Fear	2	12	68	15	14	9	56.6
Happiness	4	19	3	84	8	2	70.0
Sadness	2	8	13	5	91	1	75.8
Surprise	12	5	18	5	3	77	64.1
						Average Recognition Rate	**68.43 %**

Table 5-1 and 5-2 represents confusion matrix for audio modality using MFCC features and mel-spectrograms on BAUM-1s dataset. There were total 39 samples for Anger, 86 samples for

Disgust, 38 samples of Fear, 179 samples of Happiness, 139 samples for sadness and 43 samples of Surprise. Out of 39 samples of anger only 11 were correctly classified and other were misclassified for different emotions and the recognition rate for Anger emotion came out to be 28.2% on mel-spectrogram and 23% for MFCC . From the matrix it can be seen that system mixed up anger emotion with disgust and sadness. For Disgust only 33 samples were correctly classified on mel-spectrogram. Majority numbers of misclassified samples were of anger and sadness. For happiness most number of misclassified samples belongs to disgust and sadness. Surprise emotion was being mixed up with fear as well.

Table 5-3 and 5-4 represents confusion matrix of audio modality using MFCC and mel-spectrogramfor RML dataset. There were 120 samples in each class. From the matrix it can be seen that anger was being mixed up with disgust and sadness and fear and surprise has been mixed up together. Happiness is more likely to misclassify as disgust. A better solution to correctly classify emotions for different categories is to use multidimensional plane. In which emotion can be linked together i.e. sad smile here could be classified as sadness or disgust, but by using valence-arousal model emotion can have a better representation.

Table 5-5: Comparison with recent studies on audio emotion recognition

Dataset	**Refs**	**Recognition Rate (%)**
BAUM-1s	Zhalehpour *et al.* [41]	29.41
	Zhang *et al.* [22]	42.26
	Cornejo *et al.* [69]	46.76
	Proposed with MFCC	**43.6**
	Proposed with AlexNet	**46.23**
RML	Elmadany *et al.* [66]	58.33
	Zhang *et al.* [67]	61.86
	Cornejo *et al.* [69]	68.75
	Proposed with MFCC	**65.36**
	Proposed with AlexNet	**68.43**

Table 5-5 shows a comparison of recent studies performed in similar area on BAUM-1s and RML dataset. Zhalehpour *et al.* [41] worked with hand crafted features and got 29.41% recognition rate, Zhang *et al.* [22] worked with CNN and recognition was much better than hand crafted features used by [22]. Cornejo *et al.* [69] worked on census transform and their recognition rates were improved promisingly. We have achieved a competitive recognition rate to Cornejo *et al*. [69] for audio modality using the similar approach as them. Similarly for RML Elmadany *et al.* [66] has used traditional approach and recognition rate they achieved was 58.33. Zhang *et al.* [22] worked on CNN and achieved better recognition rates. From the comparison it can be seen that using deep learning approaches recognition rates have been improved significantly. Also working with mel-spectrogram as input for audio classification performed better than MFCC. From the experiments and comparison of recent studies it is very obvious that convolutional neural networks have won over the traditional approaches and having the data in image form performs better than raw audio signal or hand crafted audio features like MFCCs.

5.4 Visual Modality Results:

For visual modality we employed a DCNN in combination with RNN by using many to many LSTM layer. Frames were extracted from videos and by use of voile jones algorithm images were being cropped to only facial area and resized to 224x224 to fit into DCNN. The model contained 5 convolutional blocks with 2 convolutions in each block and max pooling at the end of each block. By addition of LSTM layer time factor got involved in classification process. Without LSTM each frame is judges individually and not in a sequence but with LSTM model keeps track of previous frames or sequences of video. As a video is a combination of time sequences so it is important to consider the relationship among sequences or frames. Every output in LSTM model depends upon previous outputs instead of classifying emotions individually. LOSO and LOSGO cross validation was used for experimentation of our system. Table 5-6 shows confusion matrix of speaker independent visual modality for proposed model on BAUM-1s. Table 5-7 shows confusion matrix of speaker independent visual modality for proposed model on RML dataset. Table 5-8 displays comparison of visual modality recognition rates with some recent studies on BAUM-1s and RML datasets.

Table 5-6: Confusion matrix for best results on BAUM-1s for visual modality

BAUM-1s	Anger	Disgust	Fear	Happiness	Sadness	Surprise	RR(%)
Anger	19	4	5	1	6	4	48.7
Disgust	8	45	8	5	13	7	52.3
Fear	2	5	17	2	6	7	44.7
Happiness	5	15	2	143	10	4	79.8
Sadness	7	10	5	3	100	14	71.9
Surprise	5	1	10	1	1	25	58.1
						Average Recognition Rate	**59.25 %**

Table 5-7: Confusion matrix for best results on RML for visual modality

RML	Anger	Disgust	Fear	Happiness	Sadness	Surprise	RR (%)
Anger	90	7	3	2	5	13	75.0
Disgust	10	84	7	5	11	3	70.0
Fear	8	6	80	5	7	14	66.7
Happiness	7	11	3	90	7	2	75.0
Sadness	5	9	10	5	91	0	75.8
Surprise	8	5	12	3	4	88	73.3
						Average Recognition Rate	**72.63 %**

Table 5-6 represents confusion matrix of visual modality for BAUM-1s dataset. There were total 39 samples for Anger, 86 samples for Disgust, 38 samples of Fear, 179 samples of Happiness,

139 samples for sadness and 43 samples of Surprise. Out of 39 samples of anger only 19 were correctly classified and others were misclassified for different emotions and most misclassified samples belonged to sadness. From the matrix it can be seen that system mixed up anger emotion with disgust and sadness. Disgust was mostly misclassified as sadness. And fear was mostly misclassified as surprise. Likewise, happiness was being misclassified mostly as disgust. For surprise some majority samples being misclassified after fear belonged to anger class.

Table 5-7 represents confusion matrix of visual modality for RML dataset. There were 120 samples in each class. From the matrix it can be seen that anger was being mixed up with disgust and sadness mostly but most samples were misclassified as surprise. Fear and surprise has been mixed up together. Happiness is more likely to misclassify as disgust. A better solution to correctly classify emotions for different categories is to use multidimensional plane. In which emotion can be linked together i.e. sad smile here could be classified as sadness or disgust, but by using valence-arousal model emotion can have a better representation.

Table 5-8: Comparison with recent studies on video emotion recognition

Dataset	**Refs**	**Recognition Rate (%)**
BAUM-1s	Zhalehpour *et al.* [41]	45.04
	Zhang *et al.* [44]	50.11
	Cornejo *et al.* [69]	59.52
	Proposed	**59.25**
RML	Zhang *et al.* [67]	56.90
	Elmadany *et al.* [66]	64.58
	Cornejo *et al.* [69]	75.00
	Zhang *et al.* [44]	68.09
	Proposed	**72.63**

In table 5-6 and 5-7 diagonal values shows correctly classified emotion samples. And recognition rates achieved on visual modality are 59.5% and 72.63% for BAUM-1s and RML dataset

respectively. From the results of both modalities and comparison it can be seen that our proposed methodology has achieved competitive recognition rates for both modalities.

Table 5-8 shows comparison of recent studies in the field of emotion recognition. It can be seen through the table that adding a time factor in model training has significantly improved the recognition rate especially for spontaneous reaction. However, working with different regions of face is a very good choice to improve recognition accuracy as performed by Cornejo *et al.* [69].

5.5 Fusion Results:

Two types of fusion strategies were applied in our system to combine the results of audio and video modality. Decision level and score level. Table 5.9 shows results of fusion strategies implemented. The results obtained from fusion were highly satisfactory and positive.

Table 5-9: Accuracy in % for decision level and score level fusion

Decision Level	**BAUM-1s**	**RML**
Max	56.55%	70.26%
Min	58.06%	71.32%
Sum	58.68%	72.43%
Average	58.47%	72.04%
Product	**60.64%**	**77.38%**
Score Level	59.37%	75.23%

From the results of fusion it can be seen that some fusion rules tend to decrease the accuracy of system. However, the best accuracy achieved by fusion was by using product rule as decision level. The highest value obtained by product of audio and visual score was classified as emotion label. Another fusion strategy was used in which confidence score obtained from both modalities was combined together and that passed as input feature vector to multiple classifiers SVM, random forest and KNN. Results obtained from different classifiers are shown in table 5-10. It can be seen from the results that Random Forest classifier performed better than any other classifier and shown a promising improvement in recognition rates. The accuracy of recognition

has been improved significantly after fusion of both modalities and performing late fusion on the fused data.

Table 5-10 Fusion Results with Classifier for Decision Level fusion based on product Rule

Dataset	**SVM**	**RF**	**KNN**	**LR**
BAUM-1s	60.35 %	**61.68%**	57.12%	57.13%
RML	78.07%	**79.51%**	75.64%	75.74%

Table 5-11 shows some comparison of recent studies in multimodality emotion recognition system.

Table 5-11: Comparison with recent studies on multimodal emotion recognition

Dataset	**Refs**	**Recognition Rate (%)**
BAUM-1s	Zhalehpour *et al.* [41]	51.29
	Zhang *et al.* [44]	54.57
	Cornejo *et al.* [69]	60.49
	Proposed	**61.68**
RML	Zhang *et al.* [67]	74.32
	Elmadany *et al.* [66]	75.00
	Cornejo *et al.* [69]	82.50
	Zhang *et al.* [44]	80.36
	Proposed	**79.51**

From table 5-11 it can be seen that our proposed methodology has achieved competitive recognition rates on audio-visual modality in comparison with recent studies. Zhalehpour *et al.* [41] have used early fusion and PCA for feature reduction. Zhang *et al.* [44] have used DBN for classification after fusion of audio and visual data and performed early fusion using average

pooling as their feature reduction technique. Cornejo *et.al* [69] has used early fusion and PCA and LDA was used for feature reduction. From the results and comparison it can be clearly seen that choice of fusion strategy matters a lot in improvement of recognition accuracy. We have used late fusion in our research and our results have been improved by 1.19% for BAUM-1s dataset over the state of art work. Experimental results and comparison with recent studies shows that choice of a better fusion strategy can be very crucial for achieving better recognition results.

Chapter 6

Conclusion and Future Work

CHAPTER6 : CONCLUSION AND FUTURE WORK

6.1 Conclusion:

The results show that using two modalities instead of a single one, boosts the performance of the recognition system. Both modalities may or may not offer ideal results separately but with specific fusion techniques, better recognition rates can be achieved.

We have used ALexNet model for fine tuning of audio network which is pre-trained on ImageNet dataset. Useful audio features were extracted from pre-trained AlexNet and used for audio classification. In comparison of pre-trained neural network hand crafted features like MFCC were also calculated and passed to a DCNN for audio classification. For visual recognition a deep convolutional neural network with LSTM has been used. The results show that adding time factor as a feature in training improves our classification instead of using each frame of video independently. Time factor is added in our model by using LSTM layer which holds the output until last frame of the video. Fusion of both modalities has been performed on decision level and score level. Decision level fusion with product rule gave one of the best recognition rates using random forest on employed datasets i.e. BAUM-1s (61.68 %) and RML (79.51 %). Happiness and Sadness were the most classified emotions in our datasets and fear was the most misclassified emotion. The recognition rate on BAUM-1s dataset is 61.68 % which is an improvement over previous state of art results by 1.19%.

6.2 Contributions:

- Fully automated Audio, Video and Multi-modal emotion recognition from video clips.
- First time application of CNN-LSTM on BAUM-1s dataset.
- Review and comparison of recent recognition systems designed for emotion classification.
- Detailed experiments on two different datasets using two different models for audio modality (DCNN with MFCC and Pre-trained AlexNet) and visual modality (with and without LSTM).

- Achieved one of the highest recognition rates on all two employed datasets using DCNN model. (BAUM-1s, RML).

6.3 Future Work:

The results were very competitive with some recent studies but still there is room for improvement. BAUM-1s database is very challenging due to its spontaneous nature and class imbalance. In future we plan to work on class balancing along with considering different portions of faces instead of passing whole image as input to DCNN with multiple input layers by using a non-sequential network. Only two type of fusion was implemented in our system, we also plan to work on feature level and model level fusion. Keeping that in mind, in future, more modalities like texts, postures and bio-signals can be incorporated in the system for the better understanding of emotions. Furthermore a new network called transformer has been introduced which allows parallel execution. Transformer has a non-sequential architecture and they can very useful with text like modality. And finally, new networks are introduced such as Capsule Networks which takes only few images as input for training like human being and gives very good classification. This network can also be used for emotion classification.

REFRENCES

[1] Tao, Jianhua; Tieniu Tan (2005). "Affective Computing: A Review". *Affective Computing and Intelligent Interaction*. LNCS 3784. Springer. pp. 981–995.

[2] R. W. Picard, “Affective Computing: Challenges,” Int. J. Hum. Comput. Stud., 1995.

[3] P. N. Johnson-Laird and E. Shafir, “The interaction between reasoning and decision making: an introduction,” Cognition, 1993.

[4] P. S. Bellet and M. J. Maloney, “The Importance of Empathy as an Interviewing Skill in Medicine,” JAMA J. Am. Med. Assoc., 1991.

[5] Heise, David (2004). "*Enculturating agents with expressive role behavior*". In Sabine Payr; Trappl, Robert (eds.). Agent Culture: Human-Agent Interaction in a Mutlicultural World. Lawrence Erlbaum Associates. pp. 127–142.

[6] C. E. Izard, “Basic emotions, natural kinds, emotion schemas, and a new paradigm,” Perspect. Psychol. Sci., vol. 2, no. 3, pp. 260–280, 2007.

[7] C. E. Izard, “Basic emotions, relations among emotions, and emotion-cognition relations.,” 1992.

[8] A. Ortony and T. J. Turner, “What’s basic about basic emotions?,” Psychol. Rev., vol. 97, no. 3, p. 315, 1990.

[9] P. Ekman, “An argument for basic emotions,” Cogn. Emot., vol. 6, no. 3–4, pp. 169–200, 1992.

[10] R. E. Thayer, The psychobiology of mood and arousal. Oxford University Press, Oxford, 1989.

[11] D. Bolinger and D. L. M. Bolinger, Intonation and its uses: Melody in grammar and discourse. Stanford University Press, 1989.

[12] P. Prithvi and Dr. T. Kishore Kumar, “Comparative Analysis of MFCC, LFCC, RASTA - PLP” *International Journal of Scientific Engineering and Research*, vol.4, 2016, p. 07, id: IJSER15783.

[13] Y.-L. Tian, T. Kanade, and J. F. Cohn, “Facial expression analysis,” in Handbook of face recognition, Springer, 2005, pp. 247–275.

[14] Y. Wang and L. Guan, “Recognizing human emotional state from audiovisual signals,” IEEE Trans. Multimed., vol. 10, no. 5, pp. 936–946, 2008.

[15] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, 2002.

[16] P. A. Millan Arias and J. A. Quiroga Sepulveda, "Deep Learned vs. Hand-Crafted Features for Action Classification," *2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, Laguna Hills, CA, 2018, pp. 170-171.

[17] A. Saha, S. S. Rathore, S. Sharma and D. Samanta, "Analyzing the difference between deep learning and machine learning features of EEG signal using clustering techniques," *2019 IEEE Region 10 Symposium (TENSYMP)*, Kolkata, India, 2019, pp. 660-664.

[18] Nanni, L., Ghidoni, S., & Brahnam, S. (2017). Handcrafted vs. non-handcrafted features for computer vision classification. Pattern Recognition, 71, 158–172. doi:10.1016/j.patcog.2017.05.025

[19] Bishop, C. M. *(2006), Pattern Recognition and Machine Learning, Springer*, ISBN 978-0-387-31073-2

[20] McCulloch, Warren; Walter Pitts (1943). "A Logical Calculus of Ideas Immanent in Nervous Activity". Bulletin of Mathematical Biophysics. 5 (4): 115–133. doi:10.1007/BF02478259. PMID 2185863.

[21] M. Minsky and S. A. Papert, Perceptrons: An introduction to computational geometry. MIT press, 2017.

[22] Deng, L.; Yu, D. (2014). "Deep Learning: Methods and Applications"(PDF). Foundations and Trends in Signal Processing. 7 (3–4): 1–199. doi:10.1561/2000000039.

[23] LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (28 May 2015). "Deep learning". Nature. 521 (7553): 436–444 444. Bibcode:2015Natur.521..436L. doi:10.1038/nature14539. PMID 26017442.

[24] LeCun et al., "*Backpropagation Applied to Handwritten Zip Code Recognition*," Neural Computation, 1, pp. 541–551, 1989.

[25] Ciresan, D.; Meier, U.; Schmidhuber, J. (2012). "Multi-column deep neural networks for image classification". *2012 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3642–3649. arXiv:1202.2745. doi:10.1109/cvpr.2012.6248110. ISBN 978-1-4673-1228-8.

[26] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffry (2012). "ImageNet Classification with Deep Convolutional Neural Networks" (PDF). *NIPS 2012: Neural Information Processing Systems, Lake Tahoe, Nevada*.

[27] Graves, Alex; Eck, Douglas; Beringer, Nicole; Schmidhuber, Jürgen (2003). "Biologically Plausible Speech Recognition with LSTM Neural Nets" (PDF). *1st Intl. Workshop on Biologically Inspired Approaches to Advanced Information Technology, Bio-ADIT 2004, Lausanne, Switzerland*. pp. 175–184.

[28] Santiago Fernandez, Alex Graves, and Jürgen Schmidhuber (2007). An application of recurrent neural networks to discriminative keyword spotting. Proceedings of ICANN (2), pp. 220–229.

[29] D. H. Hubel and T. N. Wiesel, “Receptive fields and functional architecture of monkey striate cortex,” J. Physiol., vol. 195, no. 1, pp. 215–243, 1968.

[30] P. J. Werbos, “Applications of advances in nonlinear sensitivity analysis,” in System modeling and optimization, Springer, 1982, pp. 762–770.

[31] S. Linnainmaa, “The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors,” Master’s Thesis (in Finnish), Univ. Helsinki, pp. 6–7, 1970.

[32] S. Linnainmaa, “Taylor expansion of the accumulated rounding error,” BIT Numer. Math., vol. 16, no. 2, pp. 146–160, 1976.

[33] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” 1985

[34] Ciresan, Dan; Ueli Meier; Jonathan Masci; Luca M. Gambardella; Jurgen Schmidhuber (2011). "*Flexible, High Performance Convolutional Neural Networks for Image Classification*" (PDF). Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence-Volume Volume Two. 2: 1237–1242. Retrieved 17 November 2013.

[35] Sergey Ioffe, and Christian Szegedy, “Batch Normalization: Accelerating Deep Network Training b y Reducing Internal Covariate Shift”, 2nd March 2015.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[37] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in Proceedings of the fourteenth international conference on artificial intelligence and statistics, 2011, pp. 315–323.

[38] O. Yadan, K. Adams, Y. Taigman, and M. Ranzato, “Multi-gpu training of convnets,” arXiv Prepr. arXiv1312.5853, 2013.

[39] Y. Wang, L. Guan, and A. N. Venetsanopoulos, “Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition,” IEEE Trans. Multimed., vol. 14, no. 3, pp. 597–607, 2012.

[40] S. Haq, T. Jan, A. Jehangir, M. Asif, A. Ali, and N. Ahmad, “Bimodal human emotion classification in the speaker-dependent scenario,” Pakistan Acad. Sci. Islam., vol. 27, 2015.

[41] M. Rashid, S. A. R. Abu-Bakar, and M. Mokji, “Human emotion recognition from videos using spatio-temporal and audio features,” Vis. Comput., vol. 29, no. 12, pp. 1269–1275, 2013.

[42] W.-Y. Chang, S.-H. Hsu, and J.-H. Chien, “FATAUVA-Net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 17–25.

[43] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, “BAUM-1: A spontaneous audio visual face database of affective and mental states,” IEEE Trans. Affect. Comput., vol. 8, no. 3, pp. 300–313, 2017.

[44] Li, H., Sun, J., Xu, Z., & Chen, L. (2017). Multimodal 2D+3D Facial Expression Recognition With Deep Fusion Convolutional Neural Network. IEEE Transactions on Multimedia, 19(12), 2816–2831. doi:10.1109/tmm.2017.2713408

[45] Kim, H.-R., Kim, Y.-S., Kim, S. J., & Lee, I.-K. (2018). Building Emotional Machines: Recognizing Image Emotions through Deep Neural Networks. IEEE Transactions on Multimedia, 1–1. doi:10.1109/tmm.2018.2827782

[46] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, “Learning Affective Features With a Hybrid Deep Model for Audio--Visual Emotion Recognition,” IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 10, pp. 3030–3043, 2018.

[47] S. Zhang, S. Zhang, T. Huang and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," in *IEEE*

Transactions on Multimedia, vol. 20, no. 6, pp. 1576-1590, June 2018, doi: 10.1109/TMM.2017.2766843.

[48] J. Zhao, X. Mao and L. Chen, "Learning deep features to recognise speech emotion using merged deep CNN," in IET Signal Processing, vol. 12, no. 6, pp. 713-721, 8 2018, doi: 10.1049/iet-spr.2017.0320.

[49] B. Yang, J. Cao, R. Ni and Y. Zhang, "Facial Expression Recognition Using Weighted Mixture Deep Neural Network Based on Double-Channel Facial Images," in IEEE Access, vol. 6, pp. 4630-4640, 2018, doi: 10.1109/ACCESS.2017.2784096.

[50] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, Dec. 2017, doi: 10.1109/JSTSP.2017.2764438.

[51] X. Xia, J. Liu, T. Yang, D. Jiang, W. Han and H. Sahli, " Video Emotion Recognition using Hand-Crafted and Deep Learning Features," 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), 2018, pp. 1-6, doi: 978-1-5386-5311-1/18

[52] H. Miao, Y. Zhang, W. Li, H. Zhang, D. Wang and S. Feng, "Chinese Multimodal Emotion Recognition in Deep and Traditional Machine Leaming Approaches," *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, Beijing, 2018, pp. 1-6, doi: 10.1109/ACIIAsia.2018.8470379.

[53] J. Zhao, S. Chen, S. Wang, Q. Jin, " Emotion Recognition using Multimodal Features" 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), 2018, pp. 1-6, doi: 978-1-5386-5311-1/18

[54] J. Y. R. Cornejo and H. Pedrini, "Audio-Visual Emotion Recognition Using a Hybrid Deep Convolutional Neural Network based on Census Transform," *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, Bari, Italy, 2019, pp. 3396-3402, doi: 10.1109/SMC.2019.8914193.

[55] I. Kansizoglou, L. Bampis and A. Gasteratos, "An Active Learning Paradigm for Online Audio-Visual Emotion Recognition," in *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2019.2961089.

[56] M. Paleari, B. Huet, and R. Chellali, “Towards multimodal emotion recognition: a new approach,” in Proceedings of the ACM international conference on image and video retrieval, 2010, pp. 174–181.

[57] K.-C. Huang, H.-Y. S. Lin, J.-C. Chan, and Y.-H. Kuo, “Learning collaborative decision-making parameters for multimodal emotion recognition,” in 2013 IEEE International Conference on Multimedia and Expo (ICME), 2013, pp. 1–6.

[58] C. Fadil, R. Alvarez, C. Martinez, J. Goddard, and H. Rufiner, “Multimodal emotion recognition using deep networks,” in VI Latin American Congress on Biomedical Engineering CLAIB 2014, Paraná, Argentina 29, 30 & 31 October 2014, 2015, pp. 813–816.

[59] D. Gharavian, M. Bejani, and M. Sheikhan, “Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks,” Multimed. Tools Appl., vol. 76, no. 2, pp. 2331–2352, 2017.

[60] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, “Audio-visual emotion recognition in video clips,” IEEE Trans. Affect. Comput., 2017.

[61] Y. Liu, Y. Li, and Y.-H. Yuan, “A Complete Canonical Correlation Analysis for Multiview Learning,” in 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 3254–3258.

[62] D. Kamińska, T. Sapiński, and G. Anbarjafari, “Efficiency of chosen speech descriptors in relation to emotion recognition,” EURASIP J. Audio Speech Music Process., vol. 2017, no. 1, p. 3, 2017.

[63] S. Haq and P. J. Jackson, “Multimodal emotion recognition,” in Machine audition: principles, algorithms and systems, IGI Global, 2011, pp. 398–423.

[64] S. Yoshizawa, N. Hayasaka, N. Wada, Y. Miyanaga, ”Cepstral gain normalization for noise robust speech recognition,” in Proc. IEEE Int. Conf. Acoust. Speech, Signal Process., Montreal, 2004, pp. 209-212.

[65] S. Zhalehpour, O. Onder, Z. Akhtar, C. E. Erdem, 'BAUM-1: A Spontaneous Audio-VÄ°sual Face Database of Affective and Mental States', IEEE Transactions on Affective Computing, Vol. 8, No.3, 2017.

[66] L. Gao, L. Qi, and L. Guan, “Information Fusion based on Kernel Entropy Component Analysis in Discriminative Canonical Correlation Space with Application to Audio Emotion Recognition,” in IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2016, pp. 2817–2821.

[67] N. E. D. Elmadany, Y. He, and L. Guan, “Multiview Learning via Deep Discriminative Canonical Correlation Analysis,” in IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2016, pp. 2409–2413

[68] “Multiview Emotion Recognition via Multi-Set Locality Preserving Canonical Correlation Analysis,” in IEEE International Symposium on Circuits and Systems. IEEE, 2016, pp. 590–593.

[69] S. Zhang, S. Zhang, T. Huang, and W. Gao, “Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition,” in ACM on International Conference on Multimedia Retrieval. ACM, 2016, pp. 281–284

[70] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Kosir, “Audio-ˇ Visual Emotion Fusion (AVEF): A Deep Efficient Weighted Approach,” Information Fusion, vol. 46, pp. 184–192, 2019.

[71] J. Cornejo and H. Pedrini, “Audio-Visual Emotion Recognition using a Hybrid Deep Convolutional Neural Network based on Census Transform,” in IEEE International Conference on Systems, Man, and Cybernetics, Bari, Italy, Oct. 2019, pp. 1–7.

[72] Mi, Aizhong, et al. “A Multiple Classifier Fusion Algorithm Using Weighted Decision Templates.” *Scientific Programming*, vol. 2016, 2016, pp. 1–10., doi:10.1155/2016/3943859.

[73] Y. Wang, L. Guan, and A. N. Venetsanopoulos, “Kernel crossmodal factor analysis for information fusion with application to bimodal emotion recognition,” IEEE Trans. Multimedia, vol. 14, no. 3, pp. 597–607, Jun. 2012.

[74] Z. Xie and L. Guan, “Multimodal information fusion of audio emotion recognition based on kernel entropy component analysis,” Int. J. Semantic Comput., vol. 7, no. 1, pp. 25–42, 2013