Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Classifying posed and natural facial expressions with the help of Deep Learning

Snehal Dey, B.Tech

A Dissertation

Presented to the University of Dublin, Trinity College
in partial fulfilment of the requirements for the degree of
Masters of Science in Computer Science (Intelligent Systems)

Supervisor: Michael Manzke

September 21, 2020

Declaration

I hereby declare that this project is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: Snehal Dey Date: 20th of September,2020

Abstract

The relaxation and contraction of facial muscles results into facial expressions. The facial muscles can be categorised into reliable and unreliable facial muscles, which have been previously successfully captured by a computer system by researchers as Kulkarni et. al., Taigman et. al.. and Hyunh et. al. This research aims at studying if (and how much better than previous studies) a Siamese Neural Network architecture, aided by Residual Network, can help in classifying the posed facial expressions as fake or genuine facial expression, keeping the result achieved by Hyunh as the baseline for comparison. SASE-FE dataset was used for this research, which is the same as Hyunh's. The system could accurately classify anger, happiness, sadness and surprise with each having an accuracy of nearly 75%, while for contempt the accuracy was 56% and the same was 61% for disgust.

Acknowledgements

I would like to thank my supervisor, Mr. Michael Manzke as his proper guidance was key to completing the work and in the process gaining critical skills for such researches. I am equally grateful to the the entire administrative and teaching staff of the computer science department under whose thoughtful handling I could manage to get past the setbacks and delays the pandemic caused. My sincere thanks to Mr. Andrew Ng as I have gained a lot of knowledge by going through his lectures regarding Neural Networks and Mr. Chayan Bhowal who cleared my doubts whenever I needed some help and clarification. Finally, I would like to thank God, as without his blessing I would not have the strength to complete my dissertation; my parents and my younger brother for being my pillars of support; and my love Abhisweta for her constant encouragement.

Contents

List of Figures

List of Tables

Nomenclature

Ann	Artificial Neural Network
Cnn	Convolutional Neural Network
Rnn	Recurrent Neural Network
ResNet	Residual Network Architecture
FACS	The Facial Action Coding System
(AUs)	Action Units
DL	Deep Learning
ROC	Receiver operating characteristics curve
AUC	Area under the curve
ReLU	Rectified Linear unit

1 Introduction

1.1 Overview

Humans are social animals and communication has always been an inevitable part of our life. Communication between humans has majorly been a two way process. Face-to-face communication (in the form of verbal and non-verbal cues) has been the earliest form of communication and till date has remained the primary source of communication. Facial expressions and human body language, i.e. body posture and gestures of hands, have served as the main sources of non-verbal communication. Out of the two non-verbal cues, the facial expression, is relatively easier to read and comprehend.

The study of human facial expression was first conducted by French researcher Duchenne De Bologne, which he published in 'Mecanisme de la Physionomie Humaine' in the year 1862 (1), where he compared the difference between a genuine human smile and a posed human smile by comparing contractions of each separate facial muscles and the wrinkles caused in the face with the help of electricity. Charles Darwin in his book 'The Expression of the Emotions in Man and Animals' (1872) praised the work of Duchenne and stated that Facial expressions are universal (2). The work done by Ekman (3) had revolutionized the way of detecting facial expressions by introducing the The Facial Action Coding System (FACS), which is by far the most used system for studying facial expression. The work done by Rinn (4) proved that sub-cortical extra-pyramidal motor system was responsible for creating the spontaneous facial expressions on the other hand voluntary facial expressions are controlled by the cortical pyramidal motor system. Thus in order to fake facial expression of emotions the pyramidal motor system plays a major role. Studies done by Hill et. al. (5) and Bartlett et. al. (6) with the help of computer vision based on dynamics of action units classified real expression of pain from false ones. However, the performance of humans in distinguishing the facial expressions in the same test done by (5) and (6) was very poor and unreliable as compared to that done by computers.

Recently, computer vision has become a very popular field with the implementation of neural networks. Computer vision is being used extensively in the field of human-computer interaction (7). Even though the usage of neural networks for finding the genuineness of emotions is quite minimal, there are few impressive works done by Kulkarni et. al. (8) and Huynch (9). Both of them had taken the help of deep learning; where Kularni et. al. had used Convolutional Neural Network for future (CNN) extraction, Huynch had used Recurrent Neural Network (RNN) for the same. Both of their works and other related work will we discussed in detail in the next section.

In this research, the main focus is on building a reliable system which helps in the classification of a genuine facial expressions of emotion from a posed facial expressions for the same emotions. The system is designed in two phases. The first phase comprises of extracting the features of facial expressions of emotion from the videos provided in the dataset using ResNet-50, which is a 50 layer deep Convolutional Neural Network (CNN) in a Siamese network. The second phase will be the usage of an eleven layer deep fully connected neural network architecture for classifying whether the facial expression of emotion detected is a genuine or a posed expression. The reason for using the deep learning approach over the traditional image processing techniques for feature extraction is due to the fact that neural networks such as CNN or RNN have dominated the domain of image classification over traditional methods ever since the work done by Krizhevsky et. al. in 2012 (10). Moreover, models like CNN in Deep Learning tend to provide greater flexibility as they have the property to be re-trained (also known as transfer learning) for use in any case with the help of a custom dataset as compared to the traditional computer vision algorithms, which are generally more domain-specific. The intuition behind the usage of Siamese network comes from the research conducted by Taigman et.al. in 2014 (11) where he introduced a better way of facial recognition from images. Hence, he came up with a system where two separate neural networks are used to extract the features from the images, where each network is given an input of an image. After the features are extracted, a similarity function such as Chi square method and triplet loss function, is used to find the similarity in the features from both of the images. If the distance between the two images is less than a certain threshold, the images are considered to be of the same person and vice versa. Siamese networks will be discussed in greater details in the background literature review along with the working of ResNet-50 architecture.

1.2 Motivation

While doing background research regarding facial expression, works done by researchers like Guillaume Duchenne (12), Charles Darwin (2) and Paul Ekman (13) state that facial expressions are universal and also that facial muscles can be grouped into two categories, i.e reliable and unreliable facial muscles. The reliable facial muscles are involuntary and hence cannot be controlled wilfully by humans in contrast to the unreliable facial muscles which are voluntary (2, 13). Therefore, by judging the contraction and relaxation of reliable facial muscles, the ingenuity of an expression can be found out. For instance, a true smile involves the contraction of obicular oculli and is often considered as a Dechunne smile (12). Thus by studying the pattern of facial muscles we can judge the genuineness of the facial expression.

Convolutional Neural Networks are now the state-of-the-art technology for computer vision related problem after the invention of AlexNet (founded by Krizhevsky, Sutskever and Hinton, 2012) (10). Previous work has been done in the same field for detection of facial expression with various methodologies, for instance: a) Kaustubh Kulkarni et. al.: Automatic recognition of facial displays of unfelt emotions (8). b) Xuan-Phung Huynh et. al.: Discrimination between genuine versus fake emotion using long-short term memory with parametric bias and facial landmarks (9). I developed an intuition that neural networks, be it CNN or RNN, are able to capture the information from the images quite well, usually known as feature extraction. Kulkarni et. al. had used a VGG-16, which is a type of CNN, for extracting information (8), while Huynh et. al. had used RNN with LSTM for feature extraction (9). After extracting the information from the images, they used boosting ensemble approach for classification of images based on boosting algorithms, which were Fisher SVM and Gradient Boosting (XGB) respectively. They both had worked on the same SASE-FE dataset. Similarly, Taigman et. al.'s work (11), 'Facial recognition via DeepFace: Closing the Gap to Human-Level Performance in Face Verification', had used a Siamese network for solving the problem of facial recognition. My current idea is based on the fact that facial expressions are universal and the question if a Siamese network architecture can help in recognizing fake and actual expression. Thus, I have come up with a methodology using ResNet50 to identify the features and then use a Siamese network type architecture to identify the true from false facial expression. However, instead of using the exact architecture as Taigman et. al., I have modified the problem to a binary classification problem using the Sigmoid activation function, i.e., I am not using a triplet loss function to find the similarity between the two images, instead I have used an eleven layer deep neural network architecture with Sigmoid activation function to detect if there is any similarity between the extracted features between the set of the images passed through the individual ResNet-50 layers and then the features are sent to the 10 layer deep neural network to classify the genuineness of the expression based on the extracted features of the facial expression. Hence, if two sets of images have same sets of features the neural network will give an output as one and if they have similarity scores below a certain threshold an output of zero will be given.

Thus, I want to find out the performance of these systems with respect to the previous work done by Xuan-Phung Huynh et. al. on discrimination between genuine versus fake emotion using long-short term memory with parametric bias and facial landmarks (9) and understand if a ResNet-50 architecture can capture features from the images, and if the eleven layers deep neural network using the sigmoid activation function can perform better than the extreme gradient boost algorithm which was used by Huynh et. al. for the classification purpose.

2 Background and Related Work

2.1 Summary of Primary Research Areas

The entire research can be broken down into sub parts which are mentioned below having a short explanation of each and its relevant importance in this research. Each of these research topics is discussed in more details in the later sections with their specific literature review. And at the end of the section we base our hypothesis based on the previous work done.

- **Facial expressions of emotions**
  This is the main area of concern for the research. Facial expression of emotion can either be genuine or posed. Hence the psychological, and physiological reasoning behind genuine and fake facial expression will be studied.

- **Neural Networks**
  Neural networks have become an inevitable part of deep learning. The only purpose behind the invention of Neural network was to simulate the human brain. In the present research, Convolution Neural Networks are used and hence the development of neural networks and its usefulness for image classification will be studied in this domain.

- **Computer vision**
  Computer vision is an interdisciplinary scientific domain where computers gets the ability to gain high level of information from digital images or videos. The present research topic comes under the domain of computer vision.

- **Related work done in the field of detection of genuineness of emotion**
  There are few works done by researchers like Kulkarni et. al., Huynh et. al. and Taigman et. al. which will be discussed in details as well as the working of the residual networks. Based on the discussion a hypothesis will be drawn for the current research.

2.2 Facial expressions of emotions

2.2.1 What are facial expressions?

Facial expressions are a result of the distortions, like contraction and relaxation of certain facial muscles simultaneously to create an expression related to a particular emotion. Based on the Ekman's theory, facial expressions can be broadly categorised into seven categories which are: i.Fear ii. Anger iii. Sadness iv. Happiness v. Surprise vi. Disgust vii. Contempt (3).

French anatomist Duchenne de Boulogne who wrote the book 'Mecanisme de la Physionomie Humaine' in 1862, was the first scientist to study human facial expression (14). However, it is considered that Charles Darwin who adopted an evolutionist approach for studying human facial expressions (15) was the first to conduct scientific study of the facial expression of emotion in 1872 (13). Charles Darwin in his work (2) has suggested that expressions of emotions through facial muscles are related to that particular emotion were innate and helped in adapting from an evolutionary point of view, and that genealogically these similarities can be observed. For example, lifting the eyebrows helped in widening the visual field to increase their vision had probably helped our ancestors respond to unexpected environmental events. Even after centuries have passed and our lifestyles have developed by many folds compared to our ancestors, the facial expression still remains in us as part of our biological characteristics and hence we still find ourselves lifting our eyebrows when something surprising happens in the environment (2, 16). The nature of the facial expressions is being communicative and they are generally used in combination with other meaningful gestures such as hand or head movements (16).

Impulses from various parts of the brain which are transmitted by the facial nucleus to the specific facial muscles causes the contraction or relaxation in the respective facial muscles. The voluntary facial expressions are a result of impulses generated from the motor cortex. In contrast to the voluntary expressions, involuntary expression is generated from the lower areas of the brain send impulses to the facial nucleus when emotions are aroused involuntarily. (13).

2.2.2 Universality of facial expressions

Individual research conducted by researchers like Floyd Allport (1924), Asch (1952) and Tomkins (1962,1963) have shown in their research that facial expressions are universal, though they had used different approaches for their hypothesis (17). These researchers have also hinted towards a possibility of the presence of facial expressions behaviour due to cultural difference and each researcher tried to come up with their own theory and explanation for this difference. A work done by Kilneberg in 1938 stated how the facial behaviours described in the Chinese literature differed from the facial behaviours with respect to a particular emotion in the Western world (18). Kilneberg had differed his view from the statement "what shows on the face is written by the culture" stating that there are certain expressive behaviours which are not based on human society or culture but are general to human beings (17). Also, a recent work done by Rachael E. Jack et. al. shows that some negative facial expressions differed among the western and the eastern people (19).

While some emotional facial expressions are explicitly used to convey message to another person such as a smiling gesture while welcoming a person or a frown to show disagreement in a discussion, there are some facial gestures which are independent of facial expressions such as a head nod. These type of expressions are quite dependent on a particular culture and vary from one person to another based on his/her cultural upbringing (17). Ekman et. al. (17), in his work for establishing the universality of facial expression had mentioned a major drawback in cross cultural research on facial expression on emotion. He said that if a common event is used to infer a common emotional state it should be kept in mind that the same event can arouse different emotions among different cultures. The observed difference in the facial expression is due to the effect of stimuli from the event and the effect of the stimuli varies from culture to culture. Thus the facial expression depicts the cultural differences more than the differences in the facial muscle associated with the stimulus of that event. For instance in one culture one might have a closed eyelid and an upturned lips during a funeral whereas people belonging from another culture might have semi closed eyelids and down turned lips. Before drawing up to conclusion that different cultures have different facial expressions for sadness it is necessary to find out what kind of emotion does the stimulus "Funeral" produce to one culture. It might be a stimulus of joy in one culture and sadness for another.

A large number of studies conducted also suggested the universality of human expressions. There is evidence that similar kinds of facial expressions can be seen when certain emotions are evoked involuntarily (20). Based on the fact that these researches have been carried out by different groups of researchers employing different techniques and methodologies for the evaluation in different laboratories involving different participants from different cultures around the world, all the experiments came out with similar kinds of results. Thus it can be accepted that the facial expressions of anger, contempt, disgust, fear, joy, sadness, and surprise are mostly universal.

2.2.3 What is the difference between genuine and fake facial expression of emotion?

A genuine facial expression is observed during the spontaneous experience of a particular emotional state, and is in correspondence with the emotional experience, while posed facial expressions are generated by an artificially stimulated unfelt emotion for the purpose of deceit or under social compliance. (3)

Duchenne tried to differentiate between the appearance of a spontaneous smile and a smile resulting from electrical stimulation of the zygomatic major muscle. In his work Duchene mentioned that when we feel joy or happiness our facial muscles respond by contracting the zygomaticus major muscle and the orbicularis oculi. Even though the movement of zygomatic major muscle is voluntary, the movement of orbicularis oculi is involuntary and cannot be moved under will. Hence while faking happiness, or performing a deceitful laughter, we can only involve the zygomaticus major muscle but not the orbicularis oculi which is involuntary. Thus, muscles around the eye (orbicularis oculi) does not obey the will; and it is activated only by a true feeling of happiness or joy. Its inertia, in smiling, unmasks a false friend. (12).

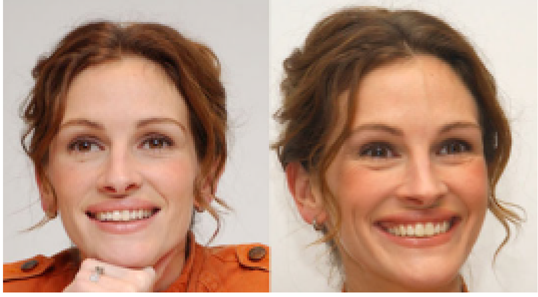


Figure 2.1: While smiling, it is observed that there is a difference between the contraction of the orbicularis oculi muscle which is involuntary around the eyes and the contraction of zygomaticus muscles which is voluntary being present in both. Left: There is an absence orbicularis oculi contraction which is regarded as a sign of an unfelt emotion or even deceitful expressions. Right: The presence of a strong orbicularis oculi contraction, creating a "crows feet" around the corners of the eyes, is considered quite often to be a sign of a genuine expressions.

The study by McLellan et.al. of neural activity was done with the help of functional magnetic resonance imaging (FMRI), taking place in the brain during display of genuine or posed facial expressions (21). It was seen that there was an increase in the activity of the brain while observing genuine compared to posed happy displays in the left medial superior frontal gyrus, the middle cingulate cortex bilaterally and the right supramarginal and left angular. It was seen overall that there was an increase in the activity for genuine displays of emotion in the left medial superior frontal and right inferior temporal gyrus. (21).

Work done by Phan et. al. in studying the brain activity related to facial expression of emotion found that the activity of basal ganglia has been related to occur in response to happy facial expressions (22), which is consistent with the rich stimulation of dopaminergic neurons that respond to positive affect (21). Keslar et. al. also discovered activity in the medial frontal cingulate sulcus region related to the displaying of genuine happy facial expression (23). While the genuine sad facial expressions has been associated with activity in the anterior cingulate (24),subcallosal cingulate (23), ventromedial prefrontal cortex (25) and temporopolar area (26).

2.2.4 Posed facial expressions

The first work on human facial expression was done by the French anatomist Duchenne de Boulogne on 1862 (27). One of the major contributions of Duchenne lies in the finding of the Duchenne smile which is regarded as a genuine smile. The Duchenne smile involves the orbicularis oculi muscle contraction which causes the cheeks to be lifted and also results in the formation of the crow's feet around the eyes together with the causing of the upward movement of the lip corners i.e zygomaticus major muscle contraction (14, 27). Charles Darwin 10 years later in 1872 adopted an evolutionist approach for studying human facial expressions (15). In his book, Charles Darwin suggested that it is difficult to deliberately impede the emotional expressions that are most difficult to voluntarily fabricate. He also suggested that it would be possible to categorise a facial expression based on the presence or absence of the difficult-to-voluntarily-generate facial actions(13). Darwin hypothesized that muscles that are difficult to voluntarily control might fail to mask a true feeling of emotion, revealing the true feelings (13, 15). "When movements, associated through habit with certain states of the mind, are partially repressed by the will, the strictly involuntary muscles, as well as those which are least under the separate control of the will, are liable still to act, and their action is often highly expressive ." (pg.54 (15)) (13) The same idea in somewhat different words: "A man when moderately angry, or even when enraged, may command the movements of his body, but those muscles of the face which are least obedient to the will, will sometimes alone betray a slight and passing emotion" (pg.79 (15)) (13). In order to study facial expressions in a more scientific way Paul Ekman and Wallace V. Friesen, who were renowned psychologists designed the concept of Facial Action Coding System (FACS) in the years 1976 and 1978. It is an anatomically based system which helps in measuring all the visible changes that in a face leading to a facial expression.

Work done by Stephen porter et. al. (28) and Paul Ekman (29) have shown that there are mainly three reasons behind the intentional manipulation of an emotional facial expression: 1. The poser is trying to stimulate an emotion without any natural feeling of that emotion. This is generally stimulating an unnatural expression. For instance when someone poses for a picture with a smile without generally feeling happy about it. 2. The poser feels a different underlying emotion but is trying to subdue the natural emotion with a different falsified expression. This is generally masking or hiding the true underlying emotion. For instance a person might not be really satisfied with a business deal but puts up a fake smile due to peer pressure. 3. The poser gets stimulated by an expression but the poser tends to keep a neutralised facial expression. For instance even though a person is angry he is trying to keep a neutral face.

2.2.5 Facial Action Coding System (FACS)

FACS is extensively used for studying human facial expressions (30). This study has been used in many fields of facial expressions such as in finding the genuineness of an expression, in detecting lies and henceforth. In the work done by Ekman based on facial muscle movement, he developed the Facial Action Coding System by identifying the presence of specific actions of 44 facial muscles called Action Units (AUs) (3, 31). Ekman in his work for studying facial expressions (32) had found out that the expression of happiness was described by the facial expressions of tensed lower eyelids, cheeks to be raised followed by the corner of the lips being pulled up. Similarly for expressions dealing with sadness, the inner eyebrows were raised and were drawn together, the corners of the lips were pulled down. For expressions relating to anger, the eyebrows were drawn together, the lower eyelids were tensed, and the lips were either pressed or parted in a shape forming a square. For expressions which conveyed fear, the eyebrows were raised and drawn together, the eyes were wide open with the lower eyelids being tensed and the lips being stretched.

An important point to be considered regarding FACS is that it is based anatomically having no 1:1 correlation with the muscle group and the AUs (3). The reason being that a single muscle might be related to different Au's by contracting in different regions producing different facial expressions. For instance when we contract the lateral portion of frontalis it raises our outer eyebrow (creating AU 2) while when we contract the medial portion it raises our inner eyebrow (creating AU 1) (3) .

2.2.6 Differentiating reliable from unreliable facial muscles

Many researchers consider emotions to be central aspects of social relationships (33, 34). Cosmides et.al. have found that behavior of a person is motivated by his emotions (35). Hence having a better control over facial expressions can be strategically used to manipulate receivers in everyday social interactions (36). Gosselin et.al. in (2010) discovered a larger number of AUs by showing that some AUs are more voluntarily controllable compared to the ones which are controllable only via co-activation by the movement of other AUs (37). Based on the work done by Duchenne (27) and Darwin (2); Ekman (13, 38) hypothesized that among the large number of facial muscles, most of the muscles are under voluntarily control and labelled them as unreliable facial muscles whereas the only few which are not readily under willful control of human were regarded as reliable facial muscle (38). He further told that the involvement of these reliable muscle for expression of emotion was a sign of the presence of specific emotion. Thus this summarizes the fact that it is quite next to impossible to control reliable facial muscles voluntarily, and the suppression of the spontaneous activation of these muscles is virtually impossible (2, 13). Thus this set of muscles is associated with emotion-specific signaling system which is based on physiological activation (36). We can summarize that Ekman's work in 2003 (13) has mainly two parts: (a)Those facial muscles which are not easily voluntarily controlled and suppress are trustworthy and reliable and (b) the activation or absence of these muscles conveys genuineness of the elicited emotion. (36) The following table depicts few reliable and versatile facial muscles.

Latin name	Name in FACS	Emotion	Controllability (M [SD]
Orbicularis oris	23:lip tighten	Anger	71.48(25.04)
Triangularis	15:lip corner depress	Sadness	68.97(28.66)
Depressor labii inferioris	16:lower lip depress	Disgust,sadness	77.9(26.65)
Frontalis, pars medialis	1:inner brow raise	Sadness	54.1(33.71)
Frontalis, pars lateralis	2:outer brow raise		
(Corrugator = AU4)	1+4	Sadness	
	1+2+4	Fear	
Risorious	20:lip stretch	Fear	78.69(24.4)
Orbicularis oculi,pars lateralis	6:cheek raise	Joy,sadness	52.24(34.01)

Table 2.1: Reliable Facial Muscles

Latin name	Name in FACS	Emotion	Controllability (M [S
Frontalis, pars medialis and lateralis	1+2:brow rise	Fear,surprise	
Corrugator	4:brow lowerer	Anger	95.72(9.63)
Orbicularis oculi ,pars palebralis	7:lid tighten	Anger	76.93(24.86)
Levator palpebrae superioris	1:upper lid raise	Anger,fear,surprise	91.62(13.8)
Zyogmaticus major	12:lip corner pull	Joy	95.24(10.46)
Levator labii superioris, caput infraorbitalis	10:upper lip raise	Sadness	83.34(20.44)
Mentalis	17:chin raise	Fear	90.83(12.19)

Table 2.2: Versatile Facial Muscles

2.2.7 Classification of a genuine from a fake expression

The work done by (28, 39) states that whenever someone tries to lie, deceive or tried to show an unfelt emotion there is always a leakage of the true genuine emotion through their nonverbal behaviour. E. D. Ross et. al. found out in his research (40) have shown that the upper part of the humans face is frequently related to the leakage of a genuine emotion, whereas the lower half of the face is said to emit cues which is manipulated most of the times for expression of an unfelt emotion. Few other researchers (41, 42) also came up with the same finding regarding the leakage genuine emotion of person while trying to suppress his natural emotion with a fake expression. A person while trying to suppress his natural emotion with a fake expression leaves behind certain cues on his/her face causing the leakage of that person's genuine emotional state either through a change in his/her pitch of voice or change in the posture or body movements of that person.(42)

This kind of false emotion is generally associated with a person who is lying or trying to deceive. Hence identifying this cues will prove to be beneficial for lie detection tests. Since the research of Ekman (13) and Darwin (2) we already know that reliable AU's are seldom controlled voluntarily. Thus examining these facial cues will help in identifying the ingenuity of an evoked facial expression of emotion. However the cues of the actual felt emotion seems to appear along with the cues related to the masked expression which makes the detection a complicated process specially for human eyes which are untrained. In order to solve such a problem image processing and machine algorithms are used to differentiate between a genuine and a posed facial expression of emotion. (8).

2.3 The Neural Networks

2.3.1 The foundation of Neural Networks

The intuition behind the development of Neural networks was to design a system that simulates the functioning of a human brain. McCulloch and Pitts in their work on 1943 (43) researched on how neurons, which are the basic cells of the brain, could produce highly complex patterns. In their work they had formalized the notion of an 'integrate and fire' neuron. The idea of associative learning in neuron was first put forward by D. O. Hebb in 1940s. He stated that — 'what fires together, wires together' (44). Frank Rosenblatt had developed Perceptron in 1957 which is the earliest implementation of a neural network. The Perceptron was a simple model consisting of input and output neurons that are capable of taking decisions based on the input vectors.However the Perceptron model was only capable of learning linear functions of the inputs.

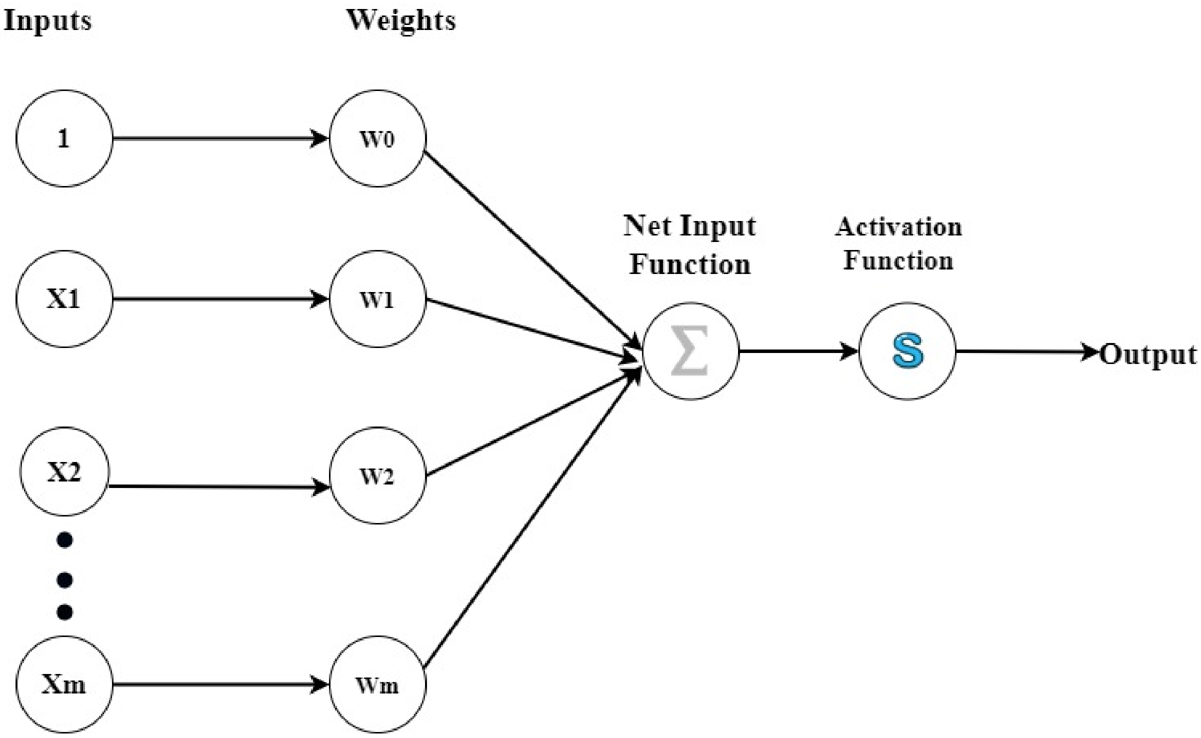


Figure 2.2: Schematic presentation of Rosenblatt's perceptron model.

In 1986 McClelland et.al. overcame the theoretical limitations of the Perceptron through the addition of hidden layers between the input and output layer and using nonlinear activation functions (45). This lead to the development of a wide range of different forms of artificial neural networks (ANNs).

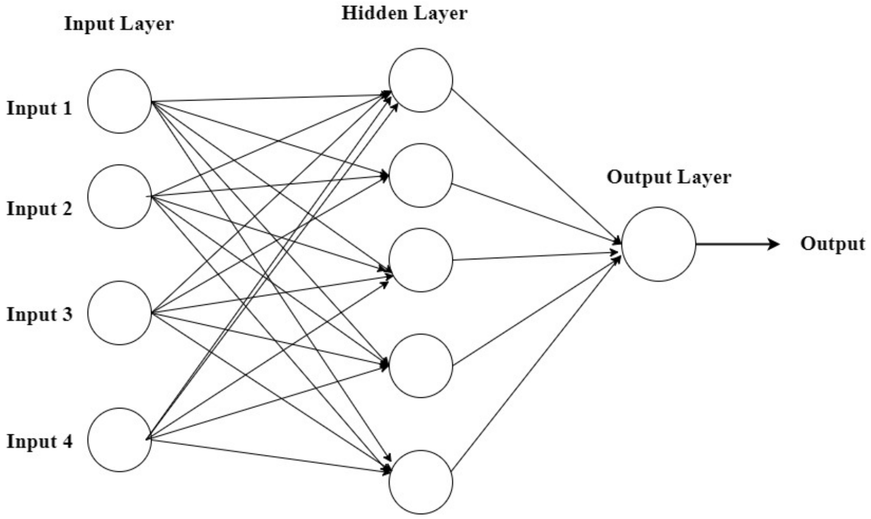


Figure 2.3: Introduction of the Hidden layer.

Although the work of McClelland had made it possible to work with functions which are non linear, it was still quite unclear regarding the training of an artificial neural network for computing any random function of interest. This lead to the development of the backward propagation algorithm in 1961, the basic concept of continuous back propagation were derived in the context of control theory by Bryson et.al. (46).The work done by Rumelhart et. al. in 1986 brought the term back-propagation into recognition and its role in neural networks (47), and it was elaborately explained by Rumelhart et. al. and gained popularity from the work (48). Although the technique had a lot of precursor models which were independently rediscovered many times from 1960s, a lot of research took place in between 1960-74 to make the backward propagation as optimal as possible (49, 50). Werbos et.al. on 1974 stated the possibility of applying this principle in an artificial neural network (51). The backward propagation algorithm provided a concrete mechanism for propagating error signals back through a multi-layer neural network causing the networks with hidden layers to be trained optimally.

The main principle behind the working of backward propagation is that it simplifies the network structure by elements weighted links that have the least effect on the trained network (51). When seen through a biological perspective back-propagation does not have a strong connection with biology. It is not really known how the neurons might propagate signals in the backward direction through multiple synapses modifying or adjusting their strengths (52). Lecun et. al. on 1989 applied back-propagation algorithm in recognizing handwritten digits taken from the U.S. Mail (53). The following table describes the development of neural networks.

Milestone/contribution	Contributor,year
MCP model, regarded as the ancestor of ANN	McCulloch and Pitts,1943
Hebbian Learning model	Hebb, 1949
First Perceptron	Rosenblatt, 1958
Back Propagation	Werbos, 1974
Neocognitron, regraded as the ancester of CNN	Fukushima 1980
Boltzmann Machine	Ackley, Hinton and Sejnowski,1985
Restricted Boltzmann Machine	Smolensky,1986
Recurrent Neural Networks	Jordan,1986
Autoencoders	Rumelhart, Hinton and Williams, 1986
LeNet, Starting the era of CNN	LeCun ,1990
LSTM	Hochreiter and Schmidhuber ,1997
Deep Belief Network	Hinton ,2006
LSTM	Hochreiter and Schmidhuber ,1997
Deep Belief Network	Hinton ,2006
Deep Boltzmann Machine	Salakhutdinov and Hinton, 2009
AlexNet, starting the age of CNN for ImageNet classification	Krizhevsky,Sutskever and Hinton, 2012
Deep Boltzmann Machine	Salakhutdinov and Hinton, 2009

Table 2.3: Development of Neural networks

2.3.2 The development of Convolutional neural network

The development of Convolutional neural network (CNN) dates back to the 1980s. Researchers like Hubel and Wiesel in their research work during the period in between 1950s and 1960s had discovered that the visual cortices of cat and monkey contain certain neurons which responded individually to small regions of the visual field (54). In 1968 Hubel and Weisel had discovered that the brain is made up of two basic categories of visual cells i.e simple cells and complex cells (55). Hubel and Wiesel had proposed a cascading model consisting of these two types of cells for pattern recognition tasks (56, 57).

Kunihiko Fukushima introduced the concept of the Neocognitron (58)in the 1980 (59, 60, 61),inspired by the previous work done by Hubel and Wiesel (55, 55) and is an extension of the cascading model proposed by them. The neocognitron is composed of multiple types of cells, S-cells and C-cells being the most important ones (58). Neocognitron had been mainly used for handwritten character recognition and other pattern recognition tasks, the convolutional neural network is said to be inspired from the neocognitron which introduced two basic types of layers in CNNs: i.convolutional layers ii. downsampling layers. (61). Atlast et.al. in 1988 used neocognitrons for the analysis of time-varying signals (62).

J. Weng et. al. introduced max-pooling an alternative to Fukushima spatial averaging in Cresceptron, maxpooling is a method where a down-sampling unit computes the most number of activations of the units in its patch (63). Modern CNNs still use this method of max-pooling(64).

Several supervised and unsupervised learning algorithms have been proposed over the decades to train the weights of a neo-cognitron (58). And currently back-propagation algorithm is being used to train the CNN architecture (61).

Alex Waibel et. al. (1987) introduced the time delay neural network (TDNN). TDNN was the first convolutional network, capable of shift-invariant classification (65). The idea behind the development of TDNN was for the phonemes classification in speech signals for automatic speech recognition, since it was challenging to automatically determine the precise segments or feature boundaries (65). The weights sharing in TDNNs takes place along the temporal dimension (66) allowing speech signals to be processed time-invariantly.

Hampshire et. al. in 1990 introduced a new model capable of performing convolutions in two dimensions (67), which outperformed all the other models for distant speech recognition (68). Yamaguchi et. al. brought up the concept of max pooling during 1990 in order to build a speaker independent isolated word recognition system by combining TDNNs with max pooling (69).

2.4 Image recognition with CNNs

Denker et. al. on 1989 designed a system for recognising ZIP Code numbers which are written by hands (70). The system consisted of convolutions kernel coefficients which were generated by laborious hand designs (53). In order to train the convolution kernel coefficients directly from images of the hand written numbers Yann LeCun et. al. in 1989 (53) applied back-propagation algorithm , making the learning process automated. It had outperformed the manual coefficient design, and was capable of performing a broader range of image recognition problems with different types of images. This approach initiated the establishment of modern computer vision.

LeCun et. al. in 1998 came up with a 7-level convolutional network called LeNet-5 (71), which was capable of classifying digits. Several banks had used LeNet-5 for recognizing numbers written on cheques. The cheques were converted into the digitized version of 32x32 pixel images. Previously in 1988 W. Zhang et al, who had come up with the TDNN for image character recognition (72, 73). W. Zhang et al in 1991 made some moderation is the architecture and training algorithm (74) and re-applied the modified architecture and training algorithm in the field of medical image processing (75)] and for diagnosing breast cancer in mammograms (76).

In the neural abstraction pyramid, the lateral and feedback connections forms the extension of the feed-forward architecture of convolutional neural networks (77). K. S. Oh and K. Jung in 2004 had proved through their work that GPUs can accelerate the standard neural networks. In their implementation they achieved a result which was 20 times faster than that of an equivalent implementation on CPU (64, 78). Although CNNs were invented in the 1980s, their success in the 2000s required fast implementations on graphics processing units (GPUs). The work of Steinkraus et. al. in 2005, also emphasised how general-purpose computing on graphics processing units (GPGPU) can accelerate the performance in machine learning (79). Chellapilla et. al. in 2006 described the implementation of a CNN with the help of GPU showing that they implemented a system on GPU (80). Their implementation which was 4 times faster than an equivalent implementation on CPU.

Ciresan et. al. in 2010 showed that GPU is faster in training deep standard neural networks with many layers by supervised learning using back propagation. In the MNIST handwritten digits benchmark their network outperformed previous machine learning methods (81). Ciresan et. al. in 2011 extended the GPU approach to CNNs. The results were amazingly impressive and they achieved an acceleration factor of 60 (82). By using such CNNs on GPU in 2011, Ciresan et. al. achieved better-than-human recognition rate of 98.98% for the first time that led them to win an image recognition contest (83).

Ciresan et. al. on 2011 had applied CNN for the task of image classification and had achieved the best result so far (2011) in the MNIST database and the NORB database (82).In 2012 Ciresan et. al. lowered down the error rate to 0.23 percent on the MNIST database (60). Alex Krizhevsky on 2012 had come up with AlexNet that won the ImageNet Large Scale Visual Recognition Challenge 2012 (10). CNNs achieved a very low error rate when it was applied for the task of facial recognition rate(84). Matsugu et.al. on 2003 had achieved a recognition rate of 97.6 percent using more than 10 subject having a total of 5,600 still images (85). The use of CNNs in the domain of video classification is relatively less as compared to those of images, due to the fact that videos are more complex than still images as videos have temporal dimension. Baccouche et.al have used CNNs can be used for video classification by extending the Convolutional Neural Networks to 3D, and learning spatio-temporal features automatically (86). Ji et.al. developed a 3D CNN architecture responsible for spawning multiple channels of information from adjoining video frames executing convolution and sub sampling. The final feature representation is obtained by combining information from all the channels (87).

2.5 Computer vision

2.5.1 What is Computer vision?

The technique through which computers get the ability to understand digital images or videos is called Computer Vision and it is an interdisciplinary scientific field. In computer vision a computer tries to understand an image or video by replicating the human visual system (88). Computer Vision serves a dual purpose. It tries to replicate the human visual system computationally when seen from a biological perspective. However if seen from an engineering perspective computer vision aims to build systems which could autonomously perform some of the tasks which the human visual system can perform and even surpass the performance of human vision in many cases (89).

2.5.2 Foundation and development of Computer vision

Larry Roberts is considered to be the father of Computer Vision whose PhD thesis led to the development of Computer vision (89). On 1963 Lawrence Roberts had come up with the machine perception of three dimensional solids (90), and this is considered to be the predecessor of computer vision. Lawrence in his work (90) tried to generate three dimensional (3D) information from a two dimensional (2D) photographs (91). In order to tackle the images from the real world, tasks such as edge detection and segmentation were needed (89). David Marr on 1978 came up with a bottom-up approach to scene understanding framework establishing that vision is hierarchical. The system's main function was in creating a 3D representations of the environment which can be interacted (92).

David G.Lowe on 1999 followed a new approach of feature based object recognition instead of trying to reconstruct objects by creating 3D models of them (93). In his work he developed a visual recognition system based on local features which were uniform to changes in rotation, location, and to an extent for illumination (93). Lowe through his research hypothesised that these features shared similar characteristics to that of the neurons involved in the process of object detection in primate vision and found in the inferior temporal cortex (93). In 2001 Paul Viola and Michael Jones (94) invented a new algorithm for real time frontal facial recognition using Haar-like features and machine learning. The Viola/Jones face detector is still widely used. It consists of a strong binary classifier built by combing weak classifiers.Then Adaboost is used to train the cascade of weak classifiers (94). Felzenszwalb et.al. on 2009 developed the Deformable Part Model (95). It decomposed objects into collections of parts, it was based on the previous work of pictorial models done by Fischler and Elschlager in the 1970s (96).The Deformable Part Model imposed a set of geometric restrictions between them, and modeled potential object centers that were taken as latent variables.

2.5.3 Usage of Deep learning in Cv for facial recognition

Face detection has been studied in computer vision since a long time and many researchers have come up with high performance face detection algorithms and recognition systems which are of great commercial interest. A variety of face recognition systems have been developed using manual features (97, 98).The feature extractor extracted the manual features from an aligned face to create a low-dimensional representation of the image, based on which a classifier made predictions. Various works such as work done by Jain et. al's Face Detection in unconstrained settings (99), Zhu et. al's .pose estimation and wilder Face- Face Detection Benchmark (100), and Yan et.al's face detection using structural models (101) have contributed to a great extent for the problems related to face detection. Broadly, face detection can be done with the help of algorithms which can be classified into mainly four categories: cascade based algorithms (102, 103), part based algorithms (101, 104),channel feature based algorithms (105, 106) and Neural networks based algorithms(107, 108).

Deep Learning (DL) which is based largely on Artificial Neural Networks (ANNs),is a subset of machine learning. However, CNNs brought about a change in the face recognition field, through their feature learning and transformation invariance properties. Before the emergence of DL, a step called feature extraction was carried out for tasks such as image classification where features are small informative patches in the images (94). Which features are to be used for the given problem has always been a problem for the traditional approach (109). As the number of classes to classify increases, feature extraction becomes more and more cumbersome (109). The recent development in the field of CNNs has tremendously influenced the field of CV in the ability to recognize objects (110).

Vaillant et. al. was on the first researchers who (111) used neural networks for face detection. A CNN was trained and in order to locate a face they scanned the image with the help of a window. Lawrence et. al, also employed CNNs in his work in the field of face recognition (84). Rowley et. al. in 1998 (112) had come up with a neural network which was connected in order to detect a face in the image. Deep learning network has become the most sought after method in Facial Expression Recognition (FER) research because of its ability to extract features powerfully. Liu et al on 2014 developed a 3DCNN model (113) capable of performing 3D convolution on image sequences with constraints based on deformable action parts. Huang et. al. in his work (114) developed a Convolutional Deep Belief Network (DBN) using Local Binary Patterns (LBP) and achieved a great performance for face verification. Barros et. al. on 2015 (115) designed a deep neural network model capable of recognizing spontaneous emotional expressions and classifying them as positive or negative. Domes et. al. on 2014 designed a facial-expression system for people suffering from Asperger Syndrome in (116).

Google's FaceNet (117) developed by Schroff et.al. and Facebook's DeepFace (11) developed by Taigman et.al use CNNs. Taigman et.al in his work (11) aligned the face to appear as a frontal face and represented the face in 3D. It is then passed down to a single convolution-pooling convolution filter, followed by three locally connected layers. The final predictions was made by the help of two fully connected layers(11, 110). Even though great performance is achieved by DeepNet, different images of the same person are not necessarily clustered while training thus its representation is not easy to interpret (110). Schroff et. al's Facenet on the other hand, made the training process learn to cluster the facial images representation of the same person by deploying a triplet loss function on the representation. OpenFace developed by Amos et. al. (118) is an open-source face recognition tool. CNNs constitute the core of OpenFace and is suitable for mobile computing, because of its smaller size and fast execution time (118).

Jung et. al. on 2015(119) combined deep temporal appearance network (DTAN) with a deep temporal geometry network (DTGN) to form a deep temporal appearance-geometry network (DTAGN) to achieve better results in facial emotion recognition. One of the branches worked in extracting the temporal appearance features from the image sequence, while the other extracted temporal geometry features from temporal facial landmarks. Ouyang et. al. on 2017 had developed (120) a new network called ResNet- Long Short Term Memory networks (LSTM) in order to get the spatio-temporal information recorded, and directly combining the lower features to LSTMs. On 2016 Vo et. al. (121) designed a model by combining global and local generic deep features which serves as an input of Support Vector Machine (SVM) for classification of the human facial expression.. Xu et. al. on 2017 (122) extracted deep features with the help of a pretrained CNN. The final presentation of facial expression was made by combining the features from the CNN together with LBPTOP features. Then, in order to classify facial expression multiclass Support Vector Machine (SVM) with one-versus-one strategy was used. Hasani et. al. on 2017 (123) combined a 3D Convolutional layers with a LSTM unit, which extracted the spatial-temporal relations between different images in the facial sequences.

Vo et. al. (121) and Xu et. al. (122) in their work extracted the features using CNN. Li et. al.(124) in his work on 2015 designed a multi-scale CNN, applying the softmax loss and regularized center loss together in order to get better performance for facial emotion recognition. Zhang et. al. (125) recently on 2019 had come up with a model capable of detecting stress on a real-time based on face detection module, the facial expression recognition module was based on the connected CNN and negative emotional stress detection module. Currently, light CNNs developed by Wu et.al. (126) and VGG Face Descriptor developed by Parkhi et.al. (127) are the most used for facial recognition.

2.6 Previous work done for detection of genuineness of the emotion

2.6.1 The work done by Huynh et al in discriminating between genuine versus fake emotion

Huynh et. al. in his work (9) tried to discriminate between genuine and the fake expressions posed by a person. Huynh, in order to discriminate the genuine expression from the posed facial expressions, had combined a mirror neuron model, i.e. RNN-PB, with a deep recurrent neural network, called long-short term memory (LSTM) with parametric bias (PB), and a binary classifier using boosting ensemble approach which is gradient boosting was used to enhance discrimination capability between the facial vectors extracted by the RNNs.

In order to conduct the research, Huynh had used the SASE-FE database. He extracted the features from the images by detecting the face in the first frame of the video, and then tracking the detected face on remaining frames of the video extracting the facial landmarks and finally learning parametric bias (PB) vectors from these images. Once the feature extraction is complete then Gradient boosting classification method is used in order to discriminate between these PB vectors. In order to detect the face in the frame, Haar-feature face detector was used (128). And in order to be robust against the problems of variation in illumination, variations in scale and posing MOSSE-based object tracker was employed(129). Face detector after detecting a region of interested (ROI) in the first frame, and then MOSSE tracks this ROI from the second frame to the final one. After the process of face detection is completed, the DLib library is used for facial landmarks detection.

The LSTM-PB operates in three modes: learning, generation, and the recognition mode.

Learning mode: In the learning mode the labeled facial expression videos, consisting of both genuine and fake emotions are provided. The goal of training is to convert the network into a time series predictor by updating weight sets to reduce the prediction error. Similarly, there is an update in the PB vectors for each training pattern to reduce the prediction error.

Generation mode: After the network has learnt, the network becomes capable of producing a stream of facial landmarks which can be either a genuine or a fake facial expression, based on the given PB vector. There is no change in the weights of the network in this mode.

Recognition mode: In the recognition mode, the prediction error between the target PB vector and the predicted PB vector is back-propagated to the PB vector based on the mean square error. If the pre-learned facial landmark movement patterns are similar to predicted PB values then the PB values tend to converge.

And finally then the LSTM-PB features are sent to the gradient boosting algorithm to classify whether the PB vectors belongs to the fake or the genuine facial expression. The gradient Boosting algorithm (130) is basically a boosting technique which involves creating a series of homogeneous weak or base learners which are single decision trees with a very little depth. In case of Adaboost algorithm, they are known as Stumps. And based on the error of each tree, the tree gets a percentage of say on the final output. After the base learners are trained, several base learners are combined to make a final strong learner for the overall prediction.

Huynh et. al. had achieved an overall result of 66.7% accuracy with an area under the curve of 0.56 for sadness, 0.64 for anger, 0.56 for happiness, 0.51 for disgust, 0.62 for surprise and 0.56 for contempt.

2.6.2 The work done by Kulkarni et. al. Automatic Recognition of Facial Displays of Unfelt Emotions

Kulkarani et. al. in his work (8) of classifying a posed expression as genuine or fake, uses a Convolutional Neural Network (CNN) to learn a static representation from still images. Then, from these features, he selected only those features which represent space along facial landmark trajectories. Using these landmark trajectories, the final features are built of varying length using a Fisher Vector encoding. And finally an SVM model is trained based on these feature vectors and a classification model is created.

In order to extract the features from the video sequences the faces were first extracted from the background, and then a pre-trained and fine tuned. VGG-16 (127) was used for recognising the facial expression. Then, they had used an emotional network knows as EMNet. The output of the VGG network was used to train the EMNet. The use of the second network is used for the computation of the static representation of the still images.

The features from the previously computed static representations were pooled along the facial landmark trajectories. Finally, these trajectories were encoded into a Fisher vector. Principal component analysis (PCA) was used in order to de-correlate the dimensions for training the Fisher vector for encoding. Then with the help of a linear SVM, the classification of these vectors takes place. Leave one out cross validation was used for all the experiments.

Kulkarni et. al. had trained their model on the SASE-FE Database and tested the model on various datasets such as CK+, Oulu-CASIA and BP4D-Spontaneous. In the CK+ dataset the author managed an impressive accuracy of 98.7% where as for the CK+ dataset they got an accuracy of 89.60%. The average F1 score for BP4D-Spontaneous dataset was 48.1%. In the Oulu-CASIA dataset the accuracy on the basis of each emotion were as follows: Anger 80.1% , Fear 95.1% , Sadness 91.3%, Disgust 88.0% , Happiness 89.7% and Surprise 92.7%.

2.6.3 Deep Residual Learning for Image Recognition

Kaiming He et al founded the Residual networks also known as the ResNet (131). It was found that considerably deeper Neural networks was difficult to train. So in order to train deeper neural networks efficiently, the authors came up with the idea of the Residual network. Based on the report (132, 133) it was found that deeper networks are prone to degradation problem, i.e as the depth of the network increases, its accuracy gets saturated followed by a rapid degradation in the accuracy. However, overfitting is not the reason behind such degradation, and higher training error is also due to the addition of more layers to a suitably deep network.

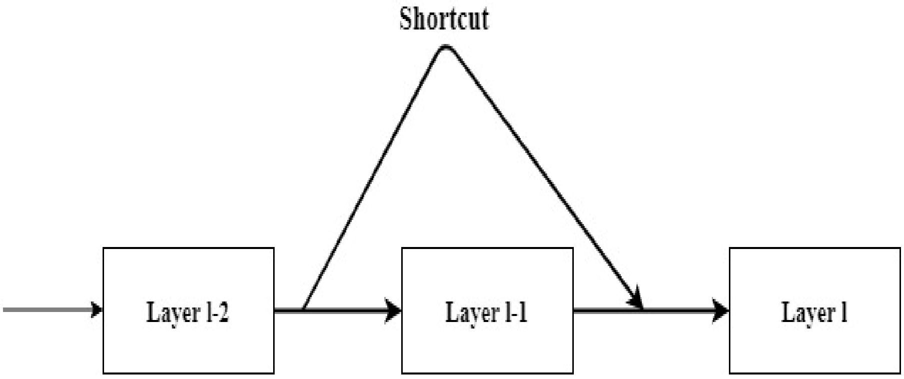


Figure 2.4: The basic structure of a ResNet architecture.

Kaiming et al had come up with the idea of designing an artificial neural network (ANN) inspired from pyramidal cells which are found in the cerebral cortex. Residual neural networks are built with the help of skip connections also known as shortcuts, which help the output from a particular layer jump over to some layers. Generally, ResNet models are implemented with Batch Normalization which help in the problem of shift variant and Rectified linear unit (Relu) activation function having double or triple layer skip connection (132). HighwayNets which is an additional weight matrix are sometimes used to learn the skip weights (133). DenseNets are models containing several parallel skips or shortcuts. A non residual network is more often described as a plain network in the context of a Residual Network(134).

The shortcut connection can be either an identity block or a convolutional block. The standard block used in ResNets is the identity block, and is used when the input activation (say a[l] a[l]) has the same dimension as the output activation (for instance a[l+2]a[l+2]).

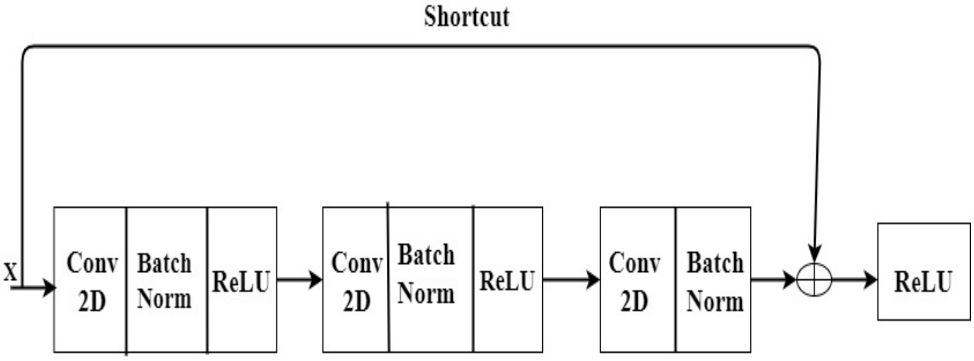


Figure 2.5: The identity block of a ResNet architecture.

On the other hand, a convolutional block is used when the input activation (say a[l] a[l]) has the different dimension as the output activation (for instance a[l+2]a[l+2]).

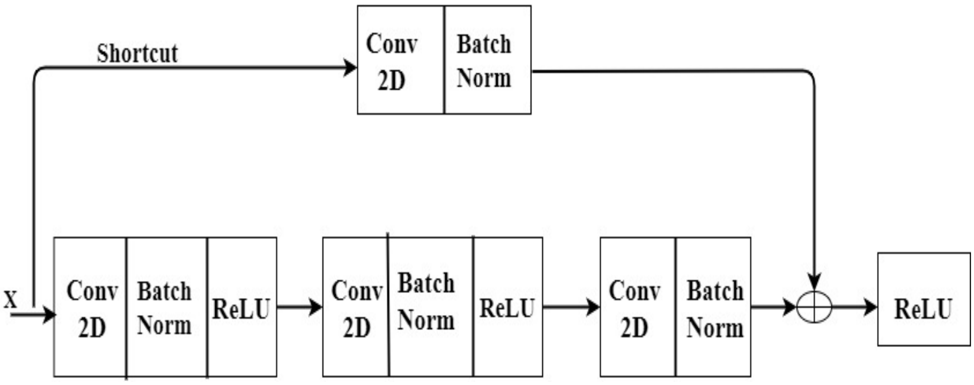


Figure 2.6: The convolutional block of a ResNet architecture.

The CONV2D layer in the shortcut path is used to resize the input X to a different dimension, so that during the final addition of the shortcut to the main path there is a dimensions match. For instance, using a 1x1 convolution with a stride of 2 can reduce the activation dimensions height and width by a factor of 2. However, non-linear activation function is not used in the CONV2D layer on the shortcut path. Its main purpose is to apply a linear function which causes the dimensions match for the later addition step by reducing the dimension of the input X.

The main intuition for skipping over layers is to avoid the problem of vanishing gradients, which causes the adjacent layers to learn its weights by reusing activations from a previous layers. Skipping over the layers simplifies the networks and it also helps in speeding up the learning process and reduced the impact of vanishing gradients.

[!h]

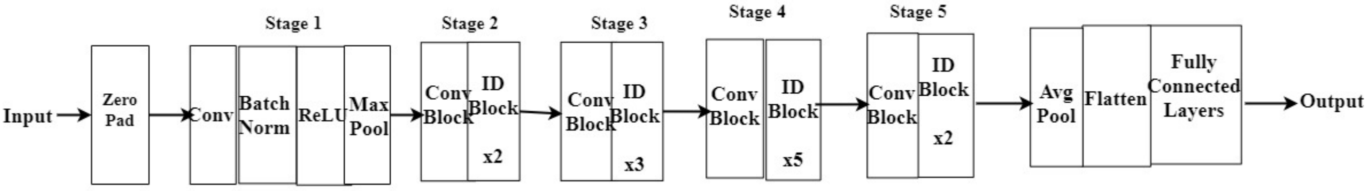


Fig 1.7: The representation of a deep residual network having 50 layers.

In order to test the performance of the ResNet architecture it was run on various datasets such as Pascal Voc, Ms Coco and its result was compared with the results of VGG-16 network architecture. For the PASCAL VOC 2007 test set, 5000 trainval images was used where as 16000 trainval images was used from the VOC 2012 for training ("07+12"). Whereas for the Pascal VOC 2012 test set, 10000 trainval and test images were used from VOC 2007 and 16000 trainval images from VOC 2012 was used for the training purpose ("07++12"). The following table shows the comparison of the results between VGG-16 and ResNet architecture. The ResNet-101 ameliorates the mean average precision (mAP) by more than 3% over the VGG-16 architecture. The increase in the mAP score is due to the fact that features learned by ResNet were more improved features as compared to VGG architecture.

Training data	07+12	07++12
Testdata	VOC07test	VOC12test
VGG16	73.2	70.4
ResNet-101	76.4	73.8

Table 2.4: Comparison of the Object detection mAP (%) value on the PASCAL VOC 2007 and 2012.

For the MS COCO dataset which has 80 object categories the PASCAL VOC metric (mAP @IoU = 0.5) and the standard COCO metric (mAP @ IoU = .5:.05:.95) was evaluated. 80,000 images were used for training the models and 40,000 images were used for validation in order to evaluate the models. ResNet-101 has been found to be better than the VGG-16 network by a 6% which is the increase in mAP@[.5, .95] over the VGG-16. There is an overall of 28% improvement, due to the fact that features learned by ResNet were more improved features as compared to VGG architecture.

Metric	mAP@.5	mAP@[.5, .95]
VGG16	41.5	21.2
ResNet-101	48.4	27.2

Table 2.5: Comparison of the object detection mAP (%) on the COCO validation set.

2.7 Deepface and Siamese networks

Bromley and LeCun et al in 1990 came up with a network which was used for signature verification which came to be known as a Siamese neural network. Siamese neural network is also known as Twin neural network is a type of ANN which shares the same parameters such as weights and bias weights between the two individual network while working simultaneously on two different input vectors to compute comparable output vectors (135, 136). Often one of the network is used to precompute the output vectors, which forms the baseline for comparison against the other output vector.

One of application of Siamese network architecture is face recognition, where one of the two network is used to precompute the images of known people and compared to an image which is given as an input to the other network to compute the similarity between the images. DeepFace which is the work done by Taigman et al (11) is an example of such a system which applies the principle of a Siamese network.

Taigman et al in his work for implementing a facial recognition used the Siamese network. After employing explicit 3D face modeling in the images the affine transformed images were sent through the Siamese network and the features were derived from each image. The features derived from each images were then used to directly evaluate whether the two input images is of the same person. This evaluation of the features from the images can be accomplished in two steps:
a) By taking the absolute difference between the features.
b) The difference is then fed to a fully connected single logistic unit. Even though the network shares the same number of parameters between the two networks,it requires twice the computation. The Siamese network's induced distance can be calculated using the formula:

$$d(f1, f2) = \sum_i \alpha_i |f1[i] - f2[i]| \tag{1}$$

where alpha are trainable parameters of the network.

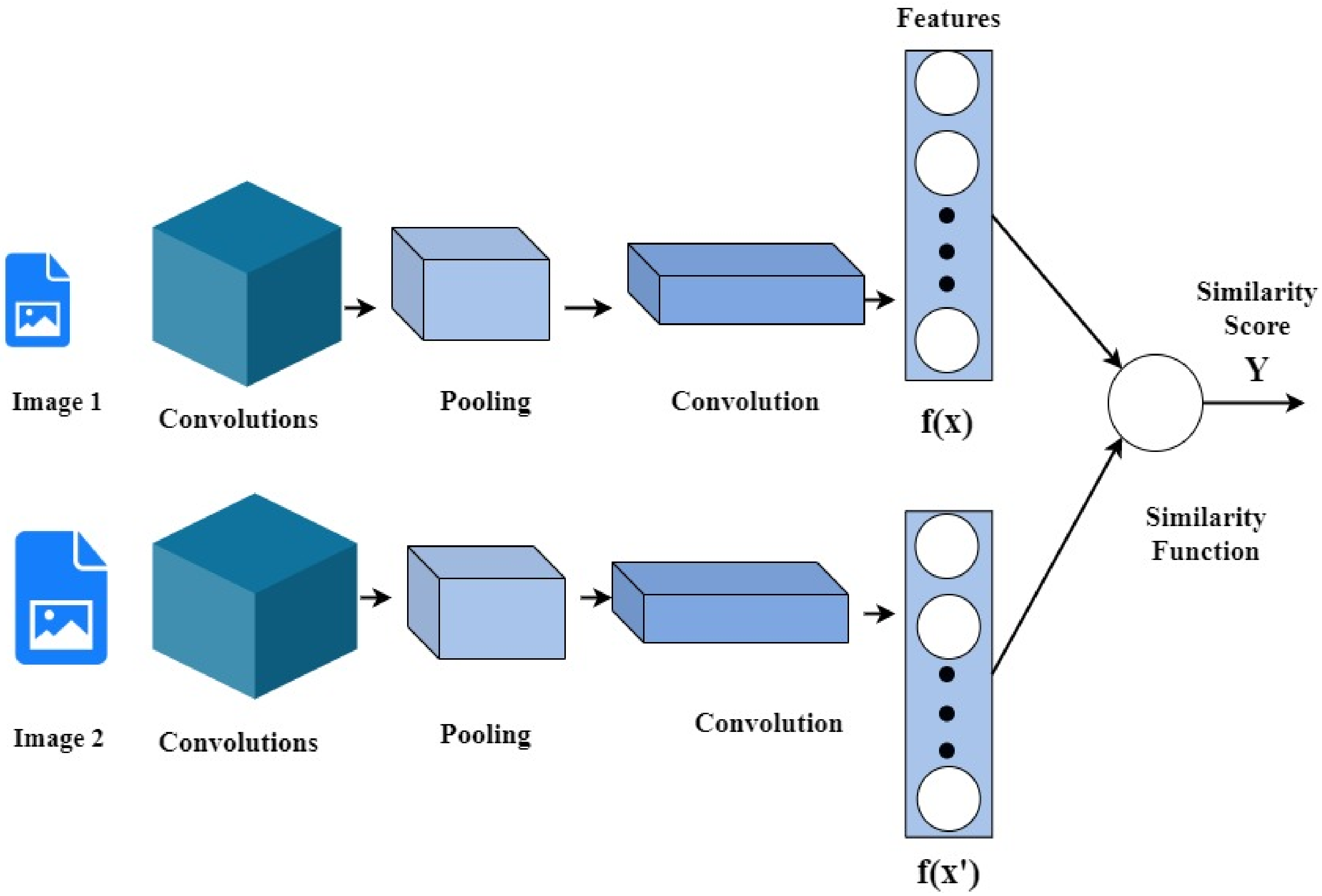


Figure 2.7: The structure of a Siamese network used in Deep face.

In the LFW dataset (137) consists photos of 5,749 celebrities with a total number of 13,323 web photos divided into 6,000 face pairs in 10 separate splits. The mean recognition accuracy is the metrics for measuring performance. It is measured using a) the restricted protocol,which consists of only the same and not same labels in training; B) the unrestricted protocol, where during training we can access additional training pairs; and C) an unsupervised setting in which the model is not trained on the the LFW images. The following table compares the accuracy of DeepFace with other works.

Method	Accuracy $\pm SE$	Protocol
Joint Bayesian (138)	0.9242 ±0.0108	restricted
Tom-vs-Pete (139)	0.9330 ±0.0128	restricted
High-dim LBP (97)	0.9517 ±0.0113	restricted
TL Joint Bayesian (98)	0.9633 ±0.0108	restricted
DeepFace-single (11)	0.9592 ±0.0029	unsupervised
DeepFace-single(11)	0.9700 ±0.0028	restricted
DeepFace-ensemble(11)	0.9715 ±0.0027	restricted
DeepFace-ensemble(11)	0.9735 ±0.0025	unrestricted

Table 2.6: Comparison of DeepFace with other models on the LFW dataset

2.8 Dataset

Facial expressions which are a result of various facial action is always associated with temporal evolution, and while interpreting an emotion displays it plays an important role (140, 141). In order to encode temporal variation so that it is easier for the recognition of subtle expressions is the main reason behind using a spatio-temporal representations. And, it is also found that spatial appearance of an expression is more prone to the identity bias than the temporal variation of an expression (142). An expression when temporally evolved can be broken down into four sub temporal segments (143): 1) neutral, 2) onset, 3) apex and 4) offset. Neutral phase consists of no muscular activity and is an expressionless phase. Onset phase denotes the starting of an expression where the facial muscle contraction starts and its intensity increases. In the Apex phase the intensity of the emotion hits a plateau and becomes stable; and during the offset the facial muscles starts relaxaing again approaching towards the Onset phase gradually. Although it is generally observed that the order of neutral-onset-apex-offset phases is usually followed, however there are also possibilities of an alternative combinations such as multiple-apex (144). In psychology the action Units (AUs) and the temporal segments are well-analysed which makes it possible to analyze sophisticated emotional states such as pain (145) and can also distinguishing between a genuine and posed facial expression (146). Facial expression of emotions are naturally perceived as a dynamic facial displays, and humans can comfortably recognize facial behaviour in video sequences as compared to the still images (147, 148). A large number of research has been carried out in assessing the ingenuity of a facial expression of emotion. However, most of the research was conducted on the basis of still images or photographs. It is also essential to keep in mind that these expressions are in itself dynamic and in order to get a better prediction accuracy it is necessary to train a model on videos, i.e a sequence of image changing with time. This sequence of images can be extracted and processed independently which will provide us with more information regarding the dynamics and changes of the facial expression with respect to time. For this particular research, mainly two types of labeled data are required, i.e having posed facial expression and a genuine facial expression.

2.8.1 The SASE-FE Database

For this research purpose, I have used the SASE-FE to train my model. The SASE-FE was created by the iCV Research Lab and contains 643 different videos captured with the help of a high resolution camera having 100 frames per second. The dataset contains video of 50 subjects with ages between 19-36 were recorded, since older adults are said to have more positive responses compared to young adults about feelings and they are quicker to regulate negative emotional states compared to the younger adults. Throughout the experiment, the participants were shown videos which are meant to induce emotional stimulus and also were asked to start with a neutral face. Each facial expression lasted for about 3-4 seconds. After illustrating each genuine facial expression, participants were asked to show a neutral face again and then the expression of a second emotion, which was the opposite of the former. For the generation of fake emotional data the participants were stimulated with a different emotional video and were asked to express a completely different emotion. For instance, a subject was shown a video which invoked the stimulus of anger and was asked to create a facial expression of surprise. Both the set of genuine and fake emotions were conducted under the same setting using the same camera. At the end of the experiment, 12 different videos of each participant were made out of which 6 were genuine facial expressions and the other six were posed facial expressions.

However, it is important to note that while preparing the SASE-FE Database factors, such as the personality or mood of the participants have been ignored for the purpose of simplicity and making the task less repetitive.

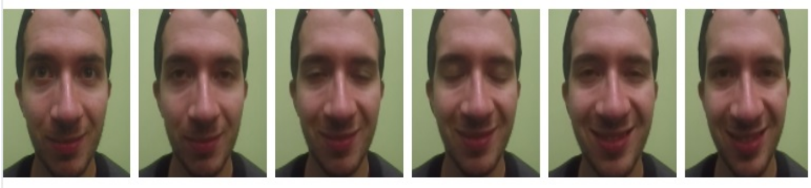


Figure 2.8: The sequence of pictures of a subject when they are naturally happy.

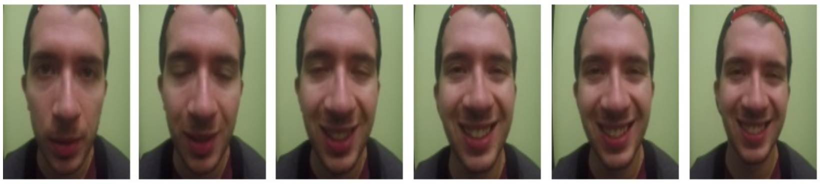


Figure 2.9: The sequence of pictures of a subject when they are faking to be happy.

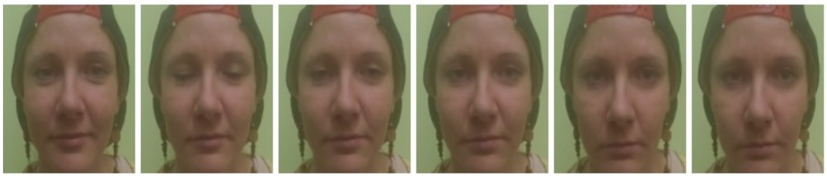


Figure 2.10: The sequence of pictures of a subject when they are naturally sad.

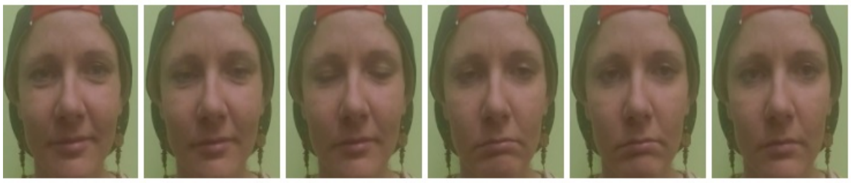


Figure 2.11: The sequence of pictures of a subject when they are faking to be sad

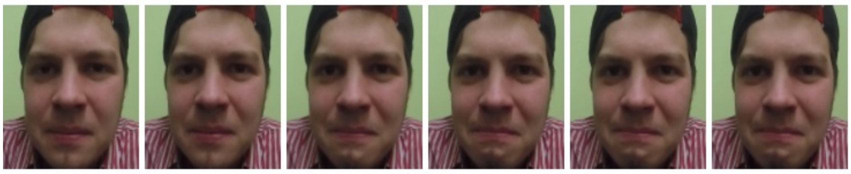


Figure 2.12: The sequence of pictures of a subject when they are naturally feeling disgusted.

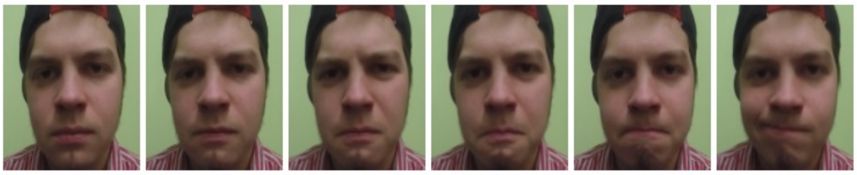


Figure 2.13: The sequence of pictures of a subject when they are faking to feel disgusted.

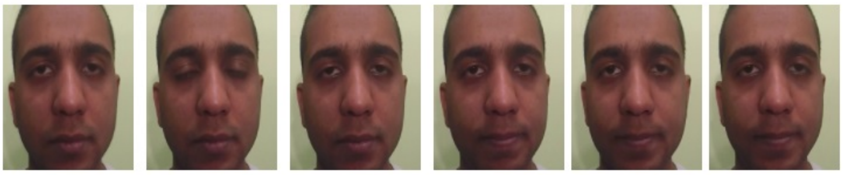


Figure 2.14: The sequence of pictures of a subject when they are naturally feeling contempt.

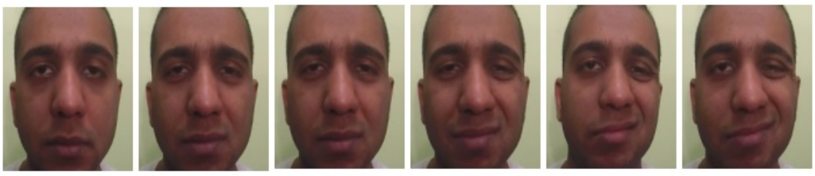


Figure 2.15: The sequence of pictures of a subject when they are faking to feel contempt.

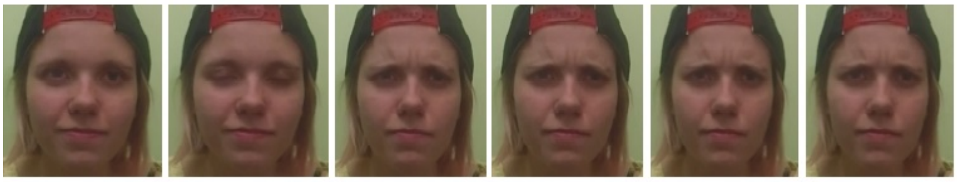


Figure 2.16: The sequence of pictures of a subject when they are naturally feeling angry.

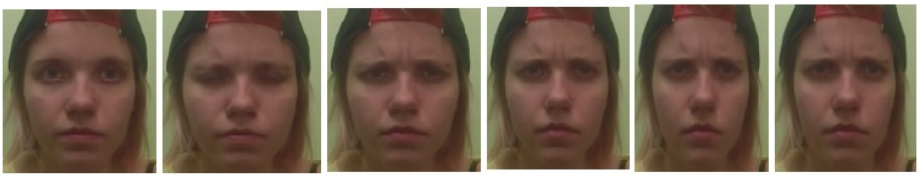


Figure 2.17: The sequence of pictures of a subject when they are faking to feel angry.

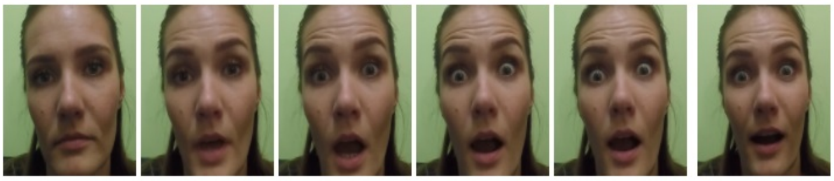


Figure 2.18: The sequence of pictures of a subject when they are naturally feeling surprised.

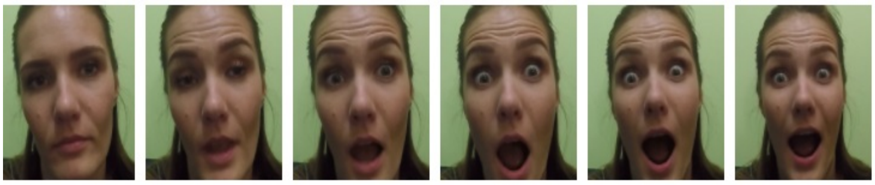


Figure 2.19: The sequence of pictures of a subject when they are faking to feel surprised.

2.9 Evaluation

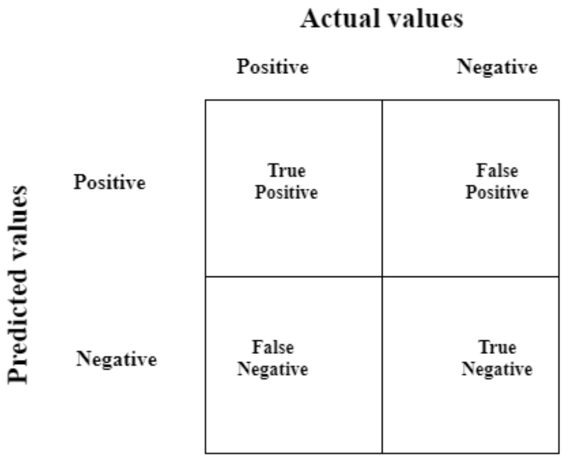
Once we have trained the model for detecting the genuineness of an expression, it is critical to validate our model on the basis of its accuracy. The confusion matrix will be the best approach to analyze the accuracy of our system on the basis of precision, recall and the F1 score.
For training the model in this particular work, the model was trained on the images of 30s. The images were obtained by extracting the frames from the videos provided in the data set. The rest of the images from the videos will be kept for validation set which will help for fine tuning the hyper parameters and a separate test set which serves for the testing the accuracy of the model.

2.9.1 Evaluation metrics

The confusion matrix is also known as error matrix (149). The confusion matrix is a tabular visualization of an algorithms performance which is usually carried out in supervised type of algorithms. For unsupervised learning it is known as matching matrix.

Confusion matrix is a kind of special kind of contingency table having two dimensions which are "actual" and "predicted". Each row in the confusion matrix represents the total instances of a particular prediction done by the algorithm where as the column represents the instances in the actual class (150).

The meaning of the terms true positive, false positive, false negative and true negative is



Predicted values	Actual values: Positive	Actual values: Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Figure 2.20: The diagram of a confusion matrix.

explained as follows:
1.True Positive (TP) : Observation is positive, and is predicted to be positive.
2.False Negative (FN) : Observation is positive, but is predicted negative.
3.True Negative (TN) : Observation is negative, and is predicted to be negative.
4.False Positive (FP) : Observation is negative, but is predicted positive.

2.9.2 Accuracy

The accuracy refers to the overall classification rate of the algorithm. which is given by the formula. It is the ratio of total correct prediction by the total prediction.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

However, accuracy should not be the sole criteria for measuring the performance of the system. Accuracy gives equal weightage to both false positive and false negative error. Thus, system having 99% accuracy can end up being a mediocre, poor or terrible depending on our problem. Thus to overcome this problem we deal with the concepts of Precision and Recall (150).

2.9.3 Precision

Precision is a ratio which helps in measuring out of all the positive classes the algorithm has predicted, how many are actually positive. To get the value of precision, we divide the total number of correctly classified positive examples by the total number of actual positive examples. High ratio of precision indicates an example labelled as positive is indeed positive, resulting in a small number of false positives (150).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

2.9.4 Recall

Recall is a ratio which helps in measuring the number of positive classes the algorithm could predict out of all the actual positive classes. It should be high as possible. To get the value of recall we divide the total number of correctly classified positive examples divide to the total number of actual positive examples. High ratio of recall indicates a small number of false negative (150).

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

When we get a high value for recall and low precision it means that even though the algorithm is able to recognize most of the positive examples having a low false negative, there are a large number of false positives which our algorithm is predicting.

On the other hand when we get a high value for precision and low value for recall, it means that even though the algorithm is missing a lot of positive examples having a high false negative, the predicted positive examples are indeed positive having a low false positive.

2.9.5 F1 Score or F-measure

Its sometimes confusing to have two metrics to determine the accuracy of the system and it is always better to combine the two measures that balances both the metrics of precision and recall by and get a single value which gives the accuracy of the algorithm. This new metric of combining the precision and recall is called F-measure or F1 score.

For calculating the value for the F-measure we take the harmonic mean of precision and recall. Harmonic Mean is used in the place of Arithmetic Mean as it penalizes the extreme values more. Thus, a high F1 score is only possible when both precision and recall have values in the same range as the F-Measure will always be nearer to the smaller value of Precision or Recall (150).

$$F - measure = \frac{2xprecisionxrecall}{precision + recall} \tag{5}$$

2.9.6 Receiver operating characteristics curve(ROC)

The ROC curve, which is the receiver operating characteristic curve, is the graphical presentation of the performance a binary classifier.In order to plot an ROC curve the true positive rate (TPR) which is also known as sensitivity or recall is plotted against the false positive rate (FPR) at various threshold settings. The false-positive rate can be calculated by (1 specificity). The ROC represents a probability density curve and the area under the curve (AUC) represents the measure of separability between two classes.The Roc and Auc gives the information that how much a classification model is capable of distinguishing between two classes i.e between a positive and negative class. The higher the AUC, the better is the model's accuracy at predicting positives as positives ane negatives as negative. Thus reducing the misclassification rates which is false positive or false negative cases (151).

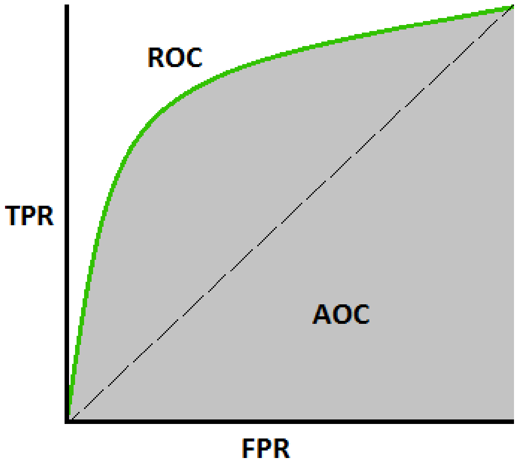


Figure 2.21: The diagram of a AUC - ROC Curve.

In a binary classification problem, the class to which a particular instance belongs i.e the prediction for each instance is based on a random continuous variable X. In case of the logistic regression these are the probabilities and in general terms is known as "score". For a threshold parameter 'T', a particular instance will belongs to a "positive" class if X>T, and "negative" otherwise. X follows a probability density $f_1(x)$ if the instance actually belongs to class "positive", and $f_0(x)$ if otherwise (151). Therefore, we can calculate the true positive rate as:

$$\mathrm{TPR}(T) = \int_T^{\infty} f_1(x)\, dx \mathrm{TPR}(T) = \int_T^{\infty} f_1(x)\, dx \tag{6}$$

The false positive rate can be calculated by:

$$\mathrm{FPR}(T) = \int_T^{\infty} f_0(x)\, dx \mathrm{FPR}(T) = \int_T^{\infty} f_0(x)\, dx \tag{7}$$

Thus the ROC curve plots the TPR(T) against the FPR(T) with T as the varying parameter parametrically. Hence the AUC - ROC curve tells us how well a given model is performing for a particular classification problem at various thresholds settings (152).

A model which can predict all the instances correctly has an AUC near of 1 which tells that it has good measure of separability. Where as a model has an AUC of 0 when it fails to predict even one instances correctly which means it has worst measure of separability. And when we get an AUC of value 0.5, it means the given model fails to separate the classes no matter what (151, 152).

2.10 Research Question

Recognizing Taigman et. al.'s work on facial recognition task (11), this research is aimed at determining whether a Siamese type of Neural Network architecture using Residual Networks can distinguish between a posed facial expression and a genuine facial expression and classify them accordingly with the help of a fully connected eleven layers deep neural network.

3 Methodology

3.1 Overview

The overall aim of this research is to investigate whether the adopted approach for detecting the genuineness of expression is capable of distinguishing the fake from real emotions with a good accuracy. This would require, for any given video which contains the facial expression which we want to judge, to break it down to frames and extract the images from the frames as 224 x 224 pixel RGB images which contain only the face of the person. Once the set of images are preprocessed, it is then passed through a Siamese network which consists of two Resnet-50 architecture for extracting the features of the images. After the feature of the images are extracted from both the ResNet, then the features are sent to a eleven layer deep fully connected neural network having the Sigmoid activation function as the output activation function of the last layer. Thus the problem of detecting the genuineness of expression is converted to a binary classification problem. The details of the methodology will be given in the details in the following subsections.

3.2 Motivation for the methodology

The main motivation for coming up with the proposed methodology is based on the previous works done by:
1. Taigman et.al. (11) who used a Saimese network to identify whether any two given images belonged to the same person or not by using triplet loss function and Chi square statistics to calculate the similarity between two given images.
2.Kulkarni et. al. (8), who had used a CNN, which is VGG-16 network for extracting the features from the images and then applied a fisher SVM in order to perform the classification purpose for distinguishing a fake emotion from a genuine emotion.
3. Huynh et. al. (9), who had applied Long Short Term Memory (LSTM) for extraction of the features from the images and then used Extreme Gradient boosting method which is an ensemble method for detecting the genuine emotions from the fake emotion. As we have worked on the same dataset, I found Hyunh's results to be a good baseline comparison for

evaluating of the results achieved from the current research.

The intuition behind using this approach to find the genuineness of an expression comes from Paul Ekman's work of Universality of Facial Expression (17) and (13). Being majorly inspired by the work done by Taigman et. al. (11) I used a Siamese network for identifying genuineness of a expression. However, I have differed from Taigman by using an eleven layer deep Neural Network instead of the similarity score and triplet loss used in the original paper. Hereafter, for extracting the features from the images, I have deployed a ResNet-50 architecture instead of Kulkarni's VGG-16 network model (8). The intention behind using ResNet-50 (131) is because it is a residual network which is better in mapping the functions from one layer to another and the network is deeper with 50 layers as compared to the VGG-16 having only 16 layers.

3.2.1 The architecture of the ResNet-50 model

For this research I have used a 50-layer deep Residual Network architecture. The architecture can be mainly divided into the identity block and the convolutional block:
1. Identity Block: The first layer in the identity block is a convolutional layer having the filters of shape (1,1) and a stride of (1,1) with a valid padding (where no padding is provided to the image). Then Batch normalization was applied to the outputs from the convolutional layer for dealing with the problems of invariant shift which makes the neural networks more robust. It also helps the gradient descent converge faster. After the batch normalization process, the 'ReLU' activation function is used followed by the output of the current layer being sent to the second convolutional layer filters of shape (3,3) and a stride of (1,1). This layer has 'same' type of padding (zeroes are used to pad the image). The outputs from these layer was then again Batch Normalised, and the ReLU activation function is used. The third layer again consists of a convolutional layer with the filters of shape (1,1) and a stride of (1,1) and having a Valid type of padding and was then batch normalized. A shortcut path was set between the input and the third layer. After the sum of the outputs between the third layer and the input image is done, the result is then sent to the ReLU activation function.

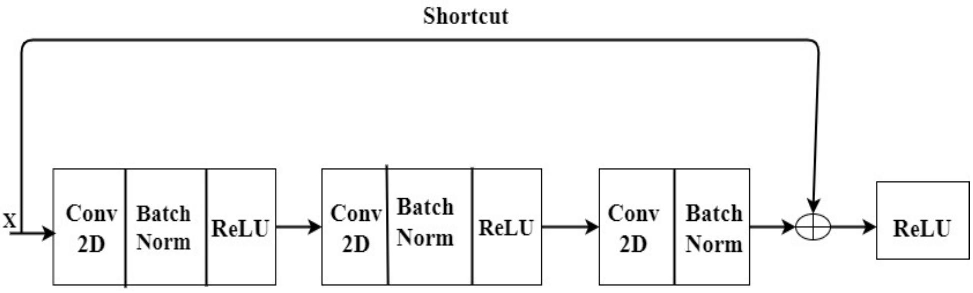


Figure 3.1: The structure of the identity block of a ResNet architecture.

2. Convolutional block: The 'First layer' in the covolutional block consists of the convolution layer, the shape of the filter is (1,1) and a stride of (2,2) and applies the 'valid' padding. Then they are Batch Normalized and then 'ReLU' activation function is used. Then the outputs of the activation layer is used by the 'Second layer' convolutional layer in the next step as the inputs. The convolutional layer had filters of shape (3,3) with a stride of (1,1) was used and 'same' type of padding was applied. Then the outputs were batch normalized and then fed to the 'ReLU' activation function is used. Then, the outputs of the activation layer are used by the convolutional layer present in the 'Third layer' as the inputs. The convolutional layer with filters of shape (1,1) with a stride of (1,1) is used and 'valid' type of padding is applied. The outputs are again Batch Normalized. A shortcut path is set between the input and the third layer. But since the dimension of the image and the outputs from the third layer are different, a covolutional layer is applied on the input image to make the dimensions match having filter of shape (1,1) and a stride of (1,1). After the outputs between the third layer and the input image are summed, the result is then to the ReLU activation function.

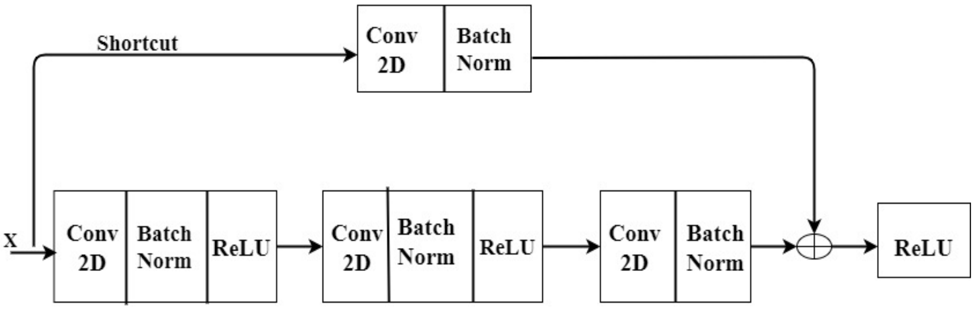


Figure 3.2: The structure of the convolutional block of a ResNet architecture.

When an image in sent through the ResNet architecture we can visualize the features that are learnt by the the ResNet architecture by plotting the feature maps using the Matplotlib library.

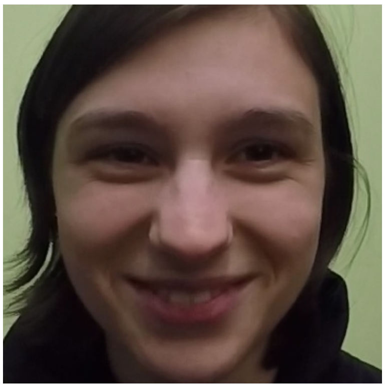


Figure 3.3: An cropped image taken from a sequence of frames of a subject showing true happiness.

After the image was passed through the ResNet architecture consisted of 50 layers, we can plot feature maps to observe what kind of features that the network is learning and we can see that initially the model take into account some abstract lines or horizontal and vertical images but as we go deeper into the model the model starts generalizing the entire face and take into account various features such as vertical and horizontal edges on the face and hence forth.

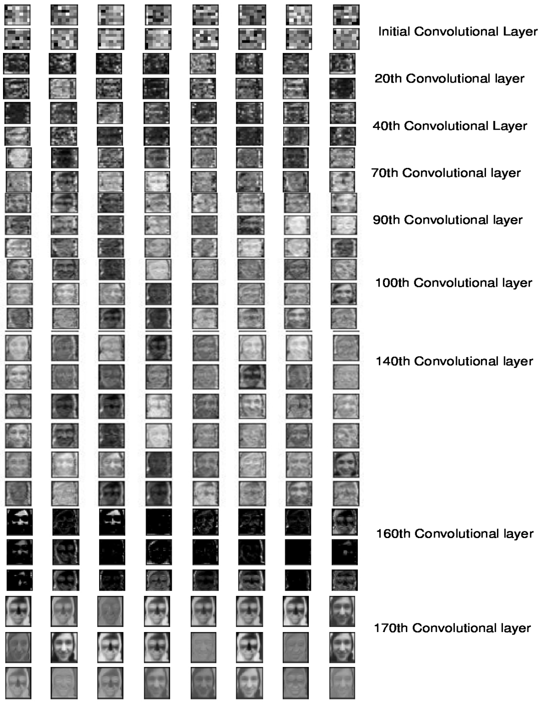

Initial Convolutional Layer
20th Convolutional layer
40th Convolutional Layer
70th Convolutional layer
90th Convolutional layer
100th Convolutional layer
140th Convolutional layer
160th Convolutional layer
170th Convolutional layer

Figure 3.4: Features learnt by a ResNet architecture.

3.2.2 Implementation of the Methodology

The implementation is divided into four stages:
1. Image extraction: This is the first part of the methodology where we extract the frames from the given videos. In order to extract the frame, opencv library (129) and MOSSE-based object tracker (153) were used. The opencv also returns a CSV file, which gives the coordinates of the region of interest (ROI) for each frame extracted, after which the frames are cropped as per the given coordinates. Since we are only dealing with facial expressions for this research, Haar-feature face detector (128) was used for extracting the faces from the frames as the region of interest. MOSSE tracker which stands for Minimum Output Sum of Squared Error is proven to be very effective and it helps in solving problems caused due to variations of illumination, changes in pose while being fast in computation (153). After the

face detector detects a ROI in the first frame, the MOSSE tracks the detected ROI from the second frame to the final frame.

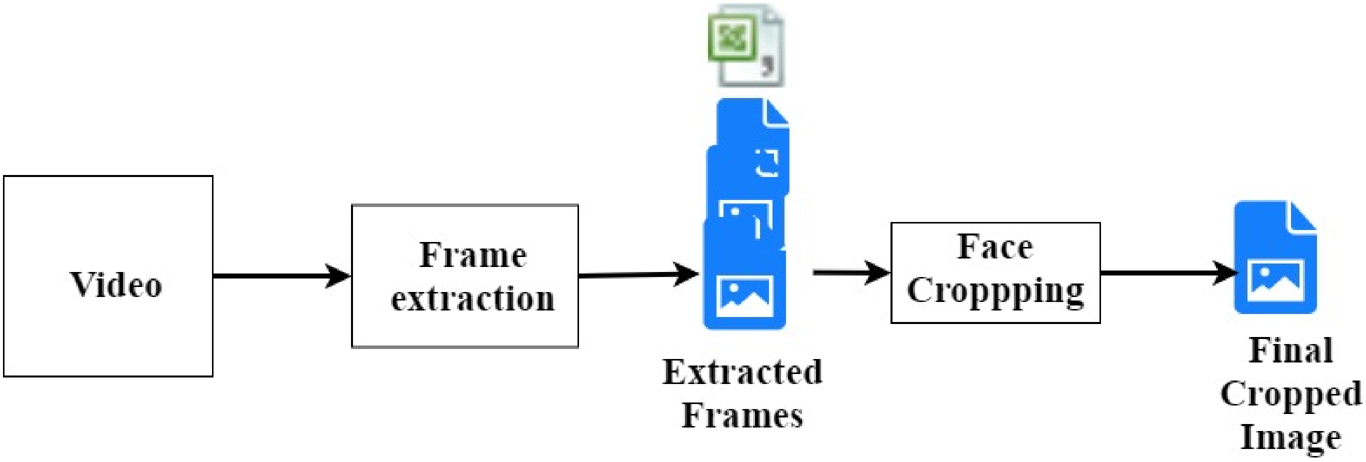


Figure 3.5: The steps involved for getting the final cropped images.

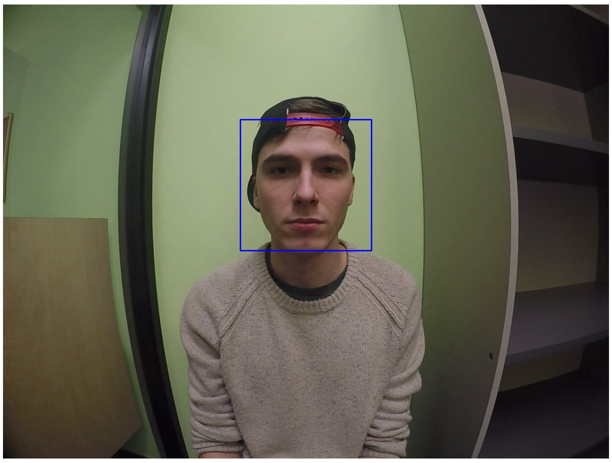


Figure 3.6: After the detection of the region of interest using Haar- feature face detector and MOOSE tracker.

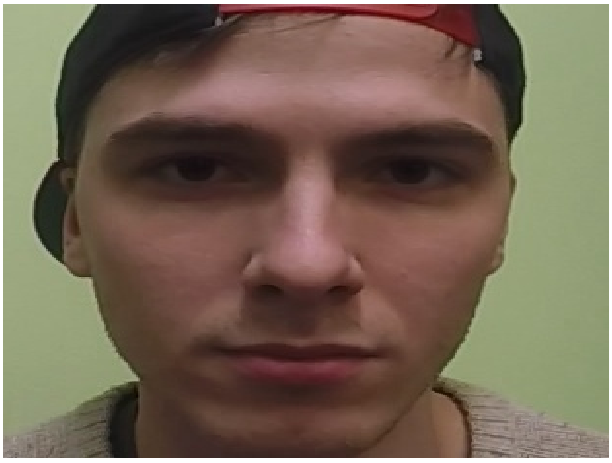


Figure 3.7: The final cropped image we get based on the coordinates of the ROI.

2. Feature extraction of the images: For this research we have used a Residual Network for extracting the features of the images. Firstly, the training data was segregated into two parts: the first part contained all the true expression images for all the training subjects and the other part contained the fake expression images. Then the images were sent through a Siamese Neural Network. To one part of the twin network all the images of the true expressions were sent while images of false expressions were sent to the other. Residual Network (ResNet) was used for feature extraction in both the twin networks. The input images were of size 224 x 224. First the images were zero padded with a pad of (3,3). Then, they were passed to the first stage where they were sent through a convolution layer having 64 filters and had the shape of (7,7) and used a stride of (2,2). Then Batch Normalization was applied in order to counteract the problem of invariant shifting as well as for the gradient descent to converge faster. After the batch normalization a max pooling layer was used with a size of (3,3) window and a stride of (2,2).

The output from the first stage was then forwarded to the second stage where first the outputs from stage 1 is fed to the the convolutional block which uses filters of size 64,64,256 respectively.

The output from the second stage was then forwarded to the third stage where first the outputs from stage 2 is fed to the convolutional block which uses filters of size 128,128,512 respectively, followed by a set of 3 identity blocks being used where each block used filters of size 128, 128, 512 respectively.

The output from the third stage was then forwarded to the fourth stage where first the outputs from stage 3 is fed to the the convolutional block uses filters of size 256, 256, 1024 respectively. And then, 5 identity blocks were applied where the identity blocks use filters of size 256, 256, 1024 respectively.

The output from the fourth stage was then forwarded to the fifth stage where first the outputs from 4th stage was fed to the convolutional block which uses filters of size 512, 512, 2048 respectively. 2 identity blocks were applied where the identity blocks use filters of size 512, 512, 2048 respectively. Finally, a 2D Average pooling layer is used which uses a window of shape (2,2) and then the outputs from the pooling layer was flattened to one vector. In order to get better features, few dense layers were added to the ResNet model. The first dense layer consisted of 4096 nodes. Batch normalization was performed for the dealing with the problem of invariant shifting and the activation function for this layer was ReLU. In order to prevent over-fitting or variance a dropout of 0.7 was used. Then, the second dense layer consisted of 1024 nodes. Batch normalization was executed and the activation function for this layer was ReLU and the layer had 0.5 as the drop out rate. Finally, the last filter having 128 was used with a ReLu activation function. The ReLu activation function was used in the last layer to get only non negative value from the final node.

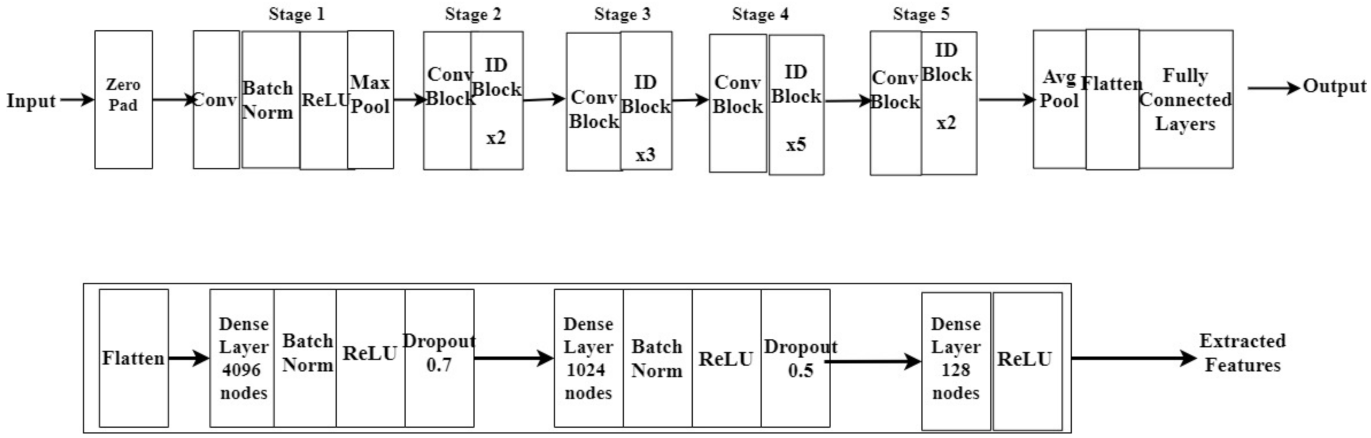


Figure 3.8: The residual network having 50 layers used in the implementation. Including the architecture of the fully connected layer.

So both the networks of the twin layer had the same architecture. In order to create the data frame from the features, the true expression images were sent to one of the paths in the twin network and a data frame of shape (62874,128) was formed. The 62874 represents each instance of a true image where each instance has 128 features. A column called results which are the labels of the features '0' was added manually to change the unlabeled dataset to a labeled dataset. Here 0 indicates that the instances belong to the class true. And the false expression images were sent to one of the paths in the twin network and a dataframe of shape (61167,128) was formed. Here 61167 represents each instances of false expression images where each instance has 128 features. A column called results, which are the labels of the features '1' was added manually to change the unlabeled dataset to a labeled dataset. Here 1 indicates that the instances belong to the class false. The two data frames after being labeled were concatenated and shuffled.
3. Training of the model for recognizing true versus false expressions: After dataframe construction, a 10 layer deep neural network was used with the sigmoid activation function at the last layer. The first layer of the neural net had 256 nodes followed by batch normalization, ReLU activation function and a drop out rate of 0.3. The second layer of the neural net had 512 nodes followed by batch normalization, ReLU activation function and a drop out rate of 0.3. The third layer of the neural net had 512 nodes again followed by batch normalization, ReLU activation function and a drop out rate of 0.3. The fourth layer of the neural net had 1024 nodes followed by batch normalization, ReLU activation function and a drop out rate of 0.4.The fifth layer of the neural net had 2048 nodes followed by batch normalization, ReLU activation function and a drop out rate of 0.5. The sixth layer of the neural net had 1024 nodes followed by batch normalization, ReLU activation function and a drop out rate of 0.3. The sixth layer of the neural net had 512 nodes followed by batch normalization, ReLU activation function and a drop out rate of 0.3. The seventh layer of the neural net had 256 nodes followed by batch normalization, ReLU activation function and a drop out rate of 0.3. The eighth layer of the neural net had 128 nodes followed by

batch normalization, ReLU activation function and a drop out rate of 0.3. Then the outputs of the activation was sent to the sigmoid activation function which gave us the probability. For gradient descent the 'Adam' optimizer was used. And binary cross entropy i.e the log loss function was used.

$$Costfunction = -1/m \sum_{m} ((results - 1) * log(results - 1) + (results) log(results)$$

Where m stands for the number of instances of output (1)

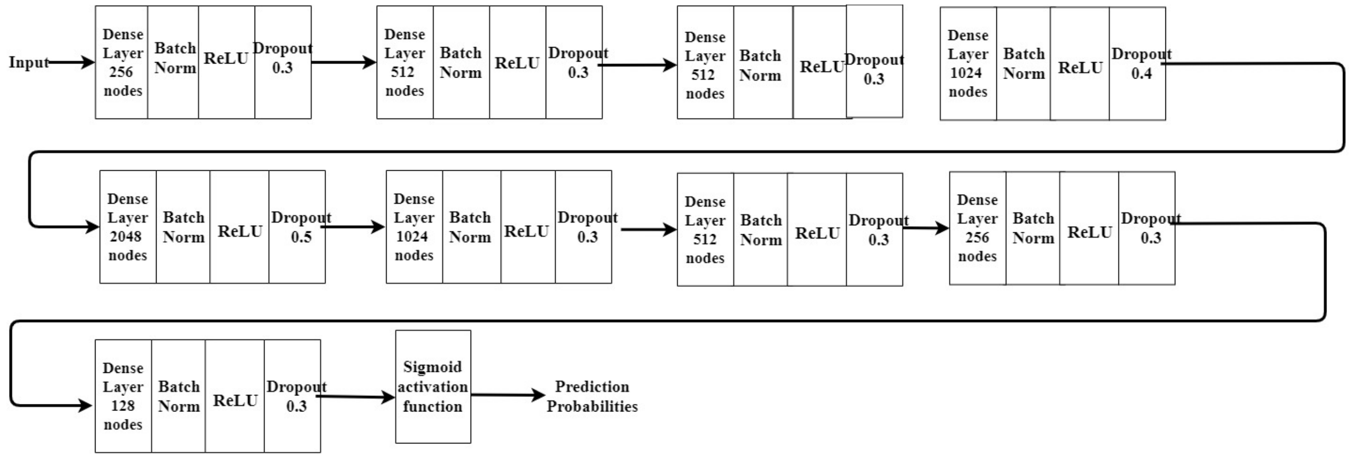


Figure 3.9: The residual network having 50 layers used in the implementation.

Thus the total model can be shown in the following picture.

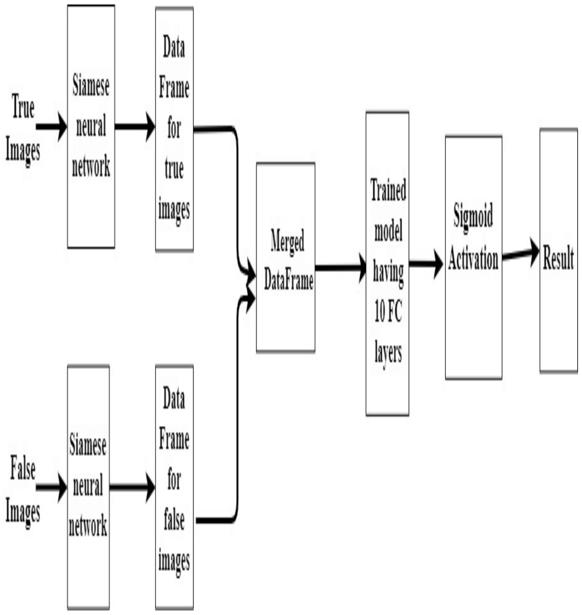


Figure 3.10: The diagram representing the total training model.

Layer (type)	Output shape	Parameter
input$_1$(*InputLayer*)	[(None, 128)]	0
fc1 (Dense)	(None, 256)	33024
bnFC1 (BatchNormalization)	(None, 256)	1024
activation (Activation)	(None, 256)	0
dropout (Dropout)	(None, 256)	0
fc2 (Dense)	(None, 512)	131584
bnFC2 (BatchNormalization)	(None, 512)	2048
activation$_1$(*Activation*)	(None, 512)	0
dropout$_1$(*Dropout*)	(None, 512)	0
fc3 (Dense)	(None, 1024)	525312
bnFC3 (BatchNormalization)	(None, 1024)	4096
activation$_2$(*Activation*)	(None, 1024)	0
dropout$_2$(*Dropout*)	(None, 1024)	0
fc4 (Dense)	(None, 1024)	1049600
bnFC4 (BatchNormalization)	(None, 1024)	4096
activation$_3$(*Activation*)	(None, 1024)	0
dropout$_3$(*Dropout*)	(None, 1024)	0
fc5 (Dense)	(None, 2048)	2099200
bnFC5 (BatchNormalization)	(None, 2048)	8192
activation$_4$(*Activation*)	(None, 2048)	0
dropout$_4$(*Dropout*)	(None, 2048)	0
fc6 (Dense)	(None, 1024)	2098176
bnFC6 (BatchNormalization)	(None, 1024)	4096
activation$_5$(*Activation*)	(None, 1024)	0
dropout$_5$(*Dropout*)	(None, 1024)	0
fc7 (Dense)	(None, 512)	524800
bnFC7 (BatchNormalization)	(None, 512)	2048
activation$_6$(*Activation*)	(None, 512)	0
dropout$_6$(*Dropout*)	(None, 512)	0
fc8 (Dense)	(None, 256)	131328
bnFC8 (BatchNormalization)	(None, 256)	1024
activation$_7$(*Activation*)	(None, 256)	0
dropout$_7$(*Dropout*)	(None, 256)	0
fc9 (Dense)	(None, 256)	65792
bnFC9 (BatchNormalization)	(None, 256)	1024
activation$_8$(*Activation*)	(None, 256)	0
dropout$_8$(*Dropout*)	(None, 256)	0
fc10 (Dense)	(None, 128)	32896
bnFC10 (BatchNormalization)	(None, 128)	512
activation$_9$(*Activation*)	(None, 128)	0
dropout$_9$(*Dropout*)	(None, 128)	0
fc11 (Dense)	(None, 1)	129

Table 3.1: Model summary of the fully connected layers

The data frame was split into training set and cross validation set having using Scikit learn library of train test split of a ratio of 70 and 30. The model was trained with a batch size of 128 and for 60 iterations. The cross validation error of the model is:

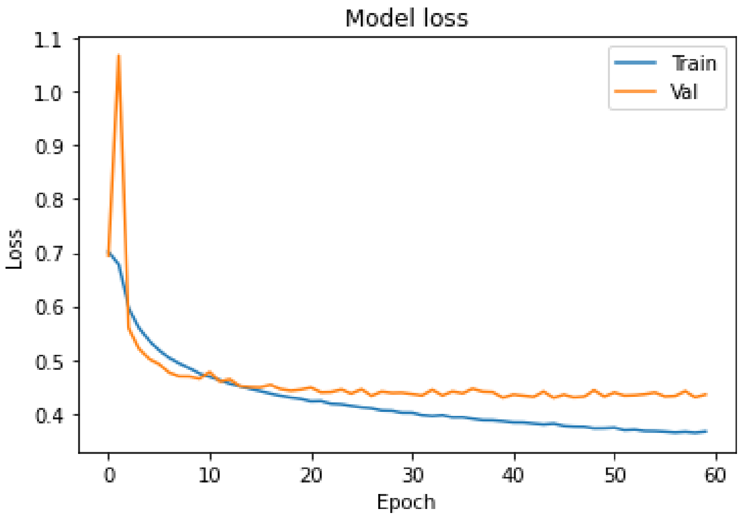


Figure 3.11: The graph showing training and cross validation error.

The cross validation set was used for hyper parameter tuning such as setting the learning rate decay of 0.9 per 10000 steps and the initial learning rate was set to 0.01. The number of layers and the droupout rates was adjusted using the cross validation set and other hyper-parameters such as β value for ADAM optimiser was unchanged. Thus, we can clearly see that the proposed model is not suffering from the problem of bias (under fitting) or variance (over fitting). Similarly the accuracy of our training model based on the cross validation set is:

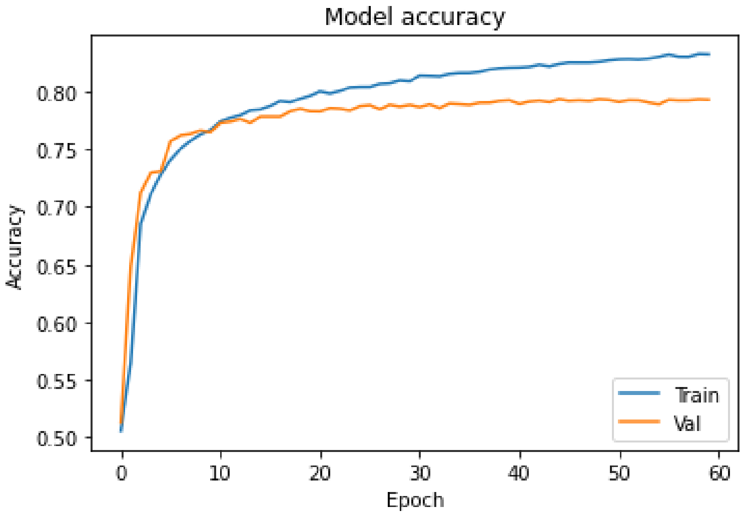


Figure 3.12: The graph showing the accuracy on training set and cross validation set

Thus our model is quite accurate and is ready to make prediction on the test data.

4. Testing the model to classify true expression from false expression: Once the model is

trained we then send the true images from the test set to the true network of the Siamese network and simultaneously use the false path for getting the features of the false images. Once the dataset is made they are concatenated shuffled and are labeled. Then they are sent to the fully connected layer and the prediction is received. The prediction is then compared with the true labels and the accuracy of the model is evaluated using the accuracy metrics. We get the Roc Auc curve for each of the six emotion.

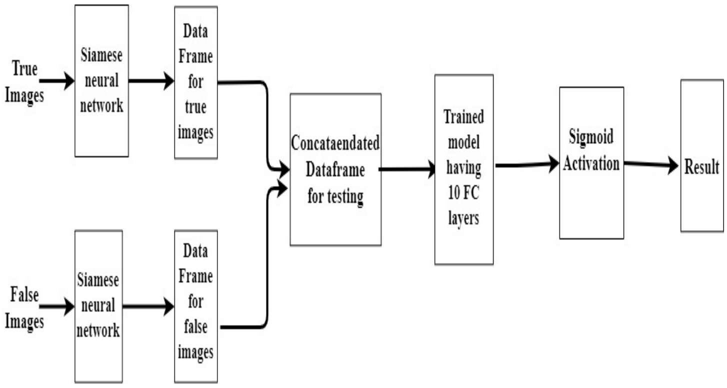
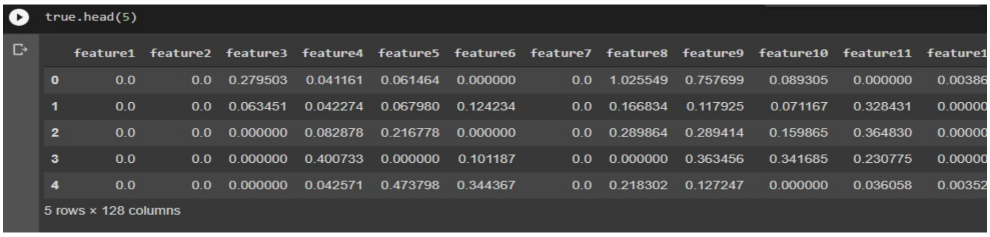

True Images
Siamese neural network
Data Frame for true images
False Images
Siamese neural network
Data Frame for false images
Concataendated Dataframe for testing
Trained model having 10 FC layers
Sigmoid Activation
Result

Figure 3.13: Testing of the pretrained model on the final dataset. FC stands for fully connected layer.

Since the image of the detailed architecture is quite large the image can be seen at (https://drive.google.com/file/d/1p44JNze8DMxPeGLm4cz65BoASwgRWO76/view?usp=sharing) this link. The fully connected layer has a total of 6,720,001 parameters out of which 6,705,921 are trainable and the remaining 14,080 were Non-trainable parameters. Whereas a single ResNet-50 architecture consists of 103,436,417 total parameters out of which the trainable parameters were: 103,373,057 and the non-trainable parameters were 63,360. Thus, a Siamese neural network in total has twice the total number of parameters both trainable ad non-trainable compared to a single ResNet-50.



```
true.head(5)

   feature1 feature2 feature3 feature4 feature5 feature6 feature7 feature8 feature9 feature10 feature11 feature1
0       0.0      0.0 0.279503 0.041161 0.061464 0.000000      0.0 1.025549 0.757699  0.089305  0.000000  0.00386
1       0.0      0.0 0.063451 0.042274 0.067980 0.124234      0.0 0.166834 0.117925  0.071167  0.328431  0.00000
2       0.0      0.0 0.000000 0.082878 0.216778 0.000000      0.0 0.289864 0.289414  0.159865  0.364830  0.00000
3       0.0      0.0 0.000000 0.400733 0.000000 0.101187      0.0 0.000000 0.363456  0.341685  0.230775  0.00000
4       0.0      0.0 0.000000 0.042571 0.473798 0.344367      0.0 0.218302 0.127247  0.000000  0.036058  0.00352
5 rows × 128 columns
```

Figure 3.14: Dataset for 5 instances of true happiness images containing 128 features for each of the instances.

3.2.3 Libraries and software used

For the research work Python was used as the programming language. The code was run on Google Colab. GPU was required for running the code. OpenCv with Haar-frontal face features were used for face detection with MOSSE filter. Keras was used for implementing the neural networks, along with the Siamese Network. Pandas were used for manipulation of the dataframes. The Matplotlib library was used for the display of graphs and metrics package was used from the Sklearn library was used for plotting the Roc Auc curve. Irfan viewer software was used to verify the coordinates of the cropped images and google drive was used to store the images and the datasets.

4 Result and Discussion

4.1 Results

After the model was trained with the data from 40 subjects among 50, the remaining 10 subjects were kept for testing of the model. Each of the ten subjects had 6 set of videos showing genuine expressions of anger, contempt, disgust, happiness, sadness and surprise, while the remaining 6 of the videos showed same six emotions that were not genuine.

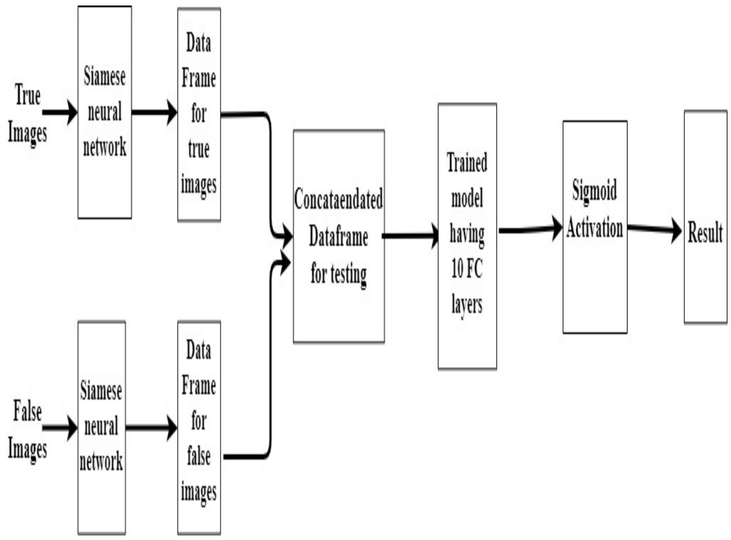

True Images
Siamese neural network
Data Frame for true images
False Images
Siamese neural network
Data Frame for false images
Concataendated Dataframe for testing
Trained model having 10 FC layers
Sigmoid Activation
Result

Figure 4.1: Testing of the pretrained model on the final dataset. FC stands for fully connected layer.

To evaluate the model, first the video frames were broken and then converted to cropped images as described in the section 3.2.2. After the images were cropped they were sent through the Siamese network. The true set of facial expression images of a particular emotion and the false set of facial expression images of that particular emotion were sent separately to the concatenated data-frame for testing. For instance, in order to evaluate how the model is performing for detection of genuineness related to happy emotion, their features were extracted with the help of ResNet-50 and a dataframe was created where each instances represented a single picture and had 128 features. The images from the true image path were given the label '0' and the images from the false images path was given a label of

'1'. This step is necessary to convert the unlabelled dataframe into a labelled dataframe. Once the dataframe was crated by labelling they were concatenated to the main dataframe and shuffled.

After that labels were extracted from the dataframe and the dataframe was passed on to the pre-trained fully-connected model. The Sigmoid activation function gave some probabilities based on the features of each instances. Once the prediction was complete, we could evaluate our model by comparing the prediction our model made with respect to the original label that we have extracted earlier from the dataset and kept it separate for evaluation of the model. The ROC curves of each of the emotion was captured. Section 2.9.6 clearly explains about ROC curve, and how Precision, Recall, Fscore value and accuracy of a model is calculated.

The ROC curve for that we got from testing the model for classifying the true expressions from the false expressions for each of the 6 emotions:

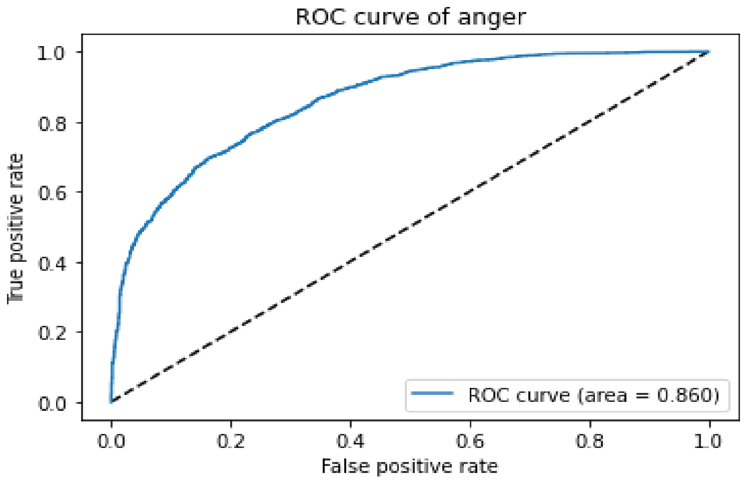


Figure 4.2: ROC curve for angry emotion.

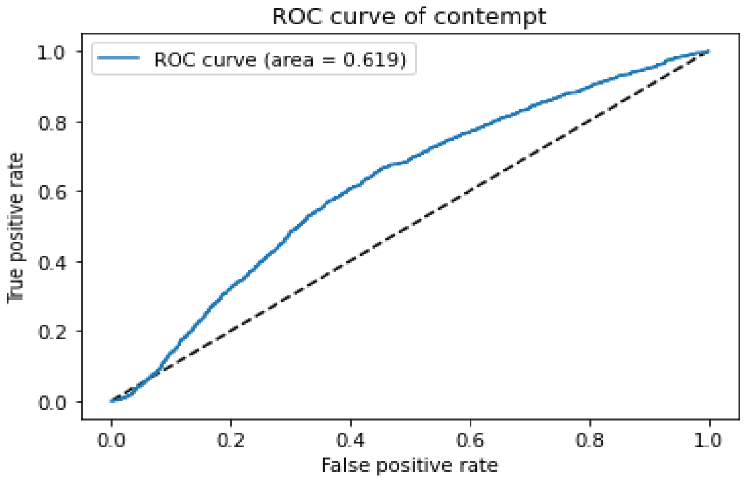


Figure 4.3: ROC curve for contempt emotion.

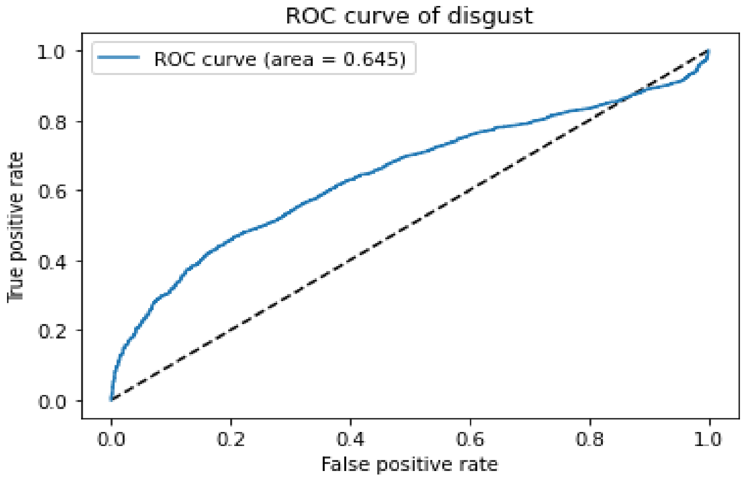


Figure 4.4: ROC curve for disgust emotion.

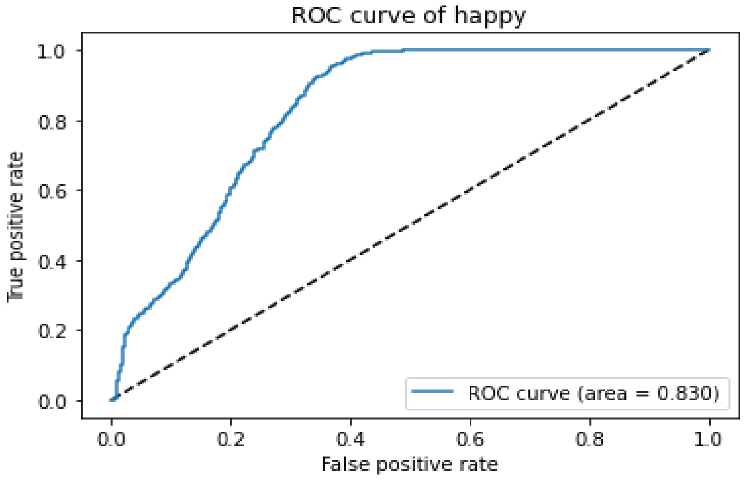


Figure 4.5: ROC curve for happiness emotion.

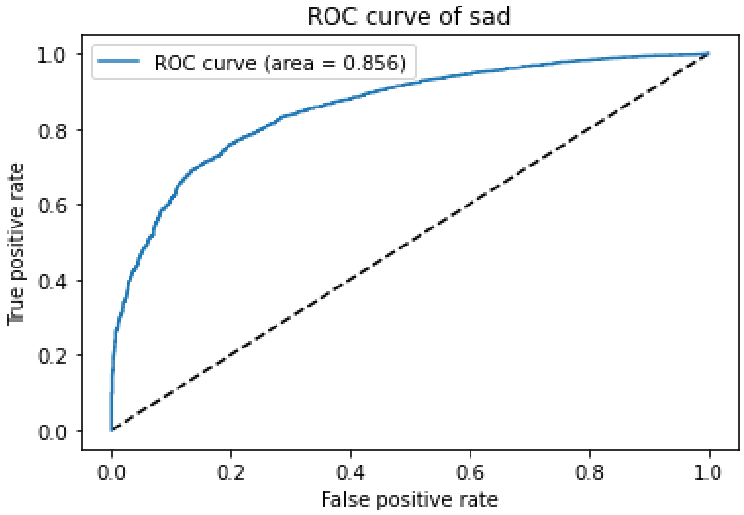


Figure 4.6: ROC curve for sadness emotion.

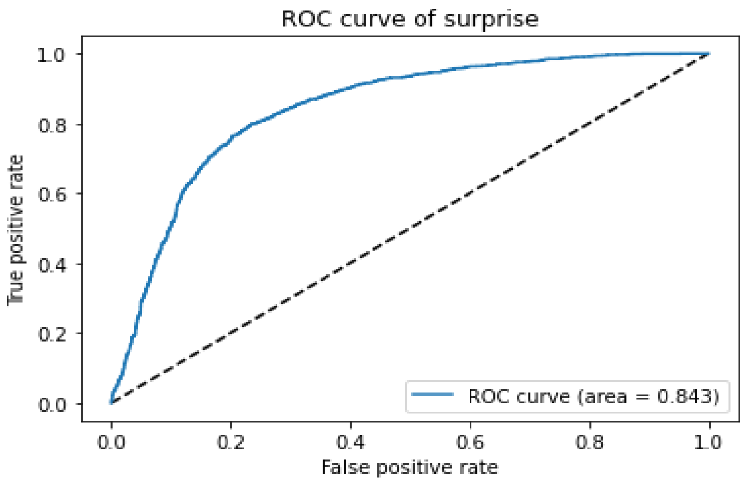


Figure 4.7: ROC curve for surprise emotion.

Emotion	Precision	Recall	F-score
Anger	0.8515878159429683	0.6629667003027245	0.745531914893617
Contempt	0.5934242181234964	0.441527446300716	0.5063291139240507
Disgust	0.6373626373626373	0.6102051551814834	0.6234883095941951
Happiness	0.7186580341377281	0.9187358916478555	0.8064729194187583
Sadness	0.7530747398297067	0.8265835929387332	0.7881188118811882
Surprise	0.6793449550977285	0.817546090273363	0.742065781881131

Table 4.1: The Precision , Recall and F-score values for each emotion

Emotion	Accuracy
Anger	0.7526881720430108
Contempt	0.5616646415552855
Disgust	0.6167943107221007
Happiness	0.7857404021937843
Sadness	0.7733651045803548
Surprise	0.7674902470741223

Table 4.2: The accuracy values for each emotion

The Roc curves were plotted with the help of Scikit learn library by importing the metrics method from the Sklearn library. From the ROC curve, we observe that for different emotions we get six different thresholds for each emotion which shall assist us in getting more true positives compared to false positives. The area under the curve tells us about the separability capability that the model holds. The Accuracy value for each emotion as well as the Precision value, Recall value and F-score value for each emotion was found out with the help of sklearn library using the metrics method.

The confusion matrix for all six emotions respectively where 0 stands for images of true emotional facial expression and 1 stands for false emotional facial expression:

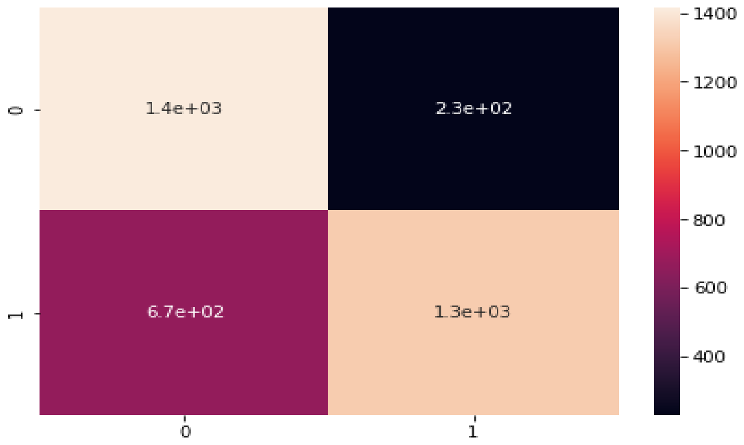


Figure 4.8: Confusion matrix for angry emotion.

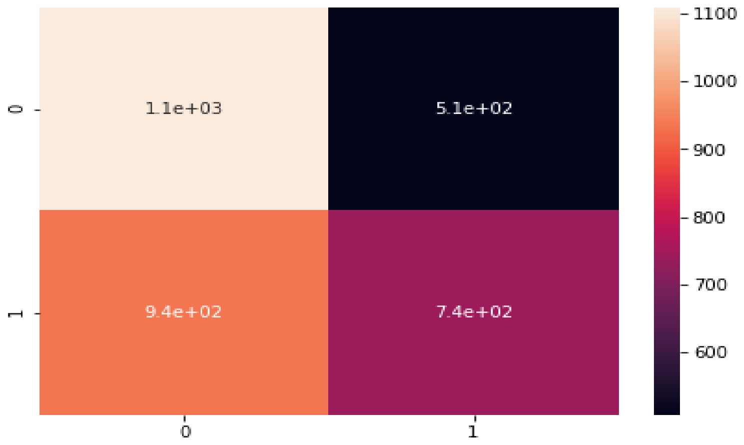


Figure 4.9: Confusion matrix for contempt emotion.

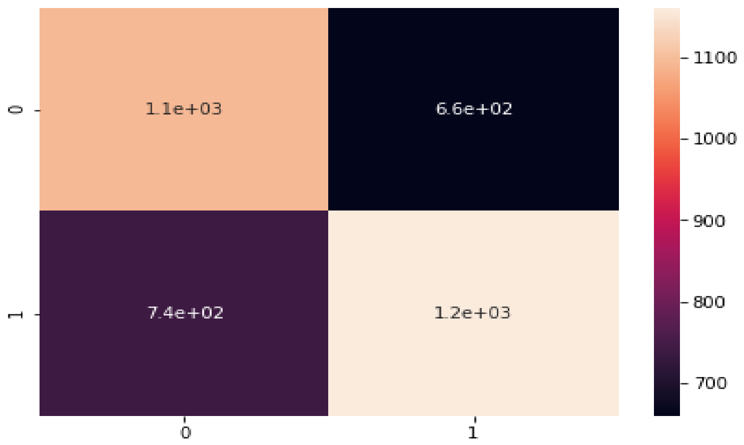


Figure 4.10: Confusion matrix for disgust emotion.

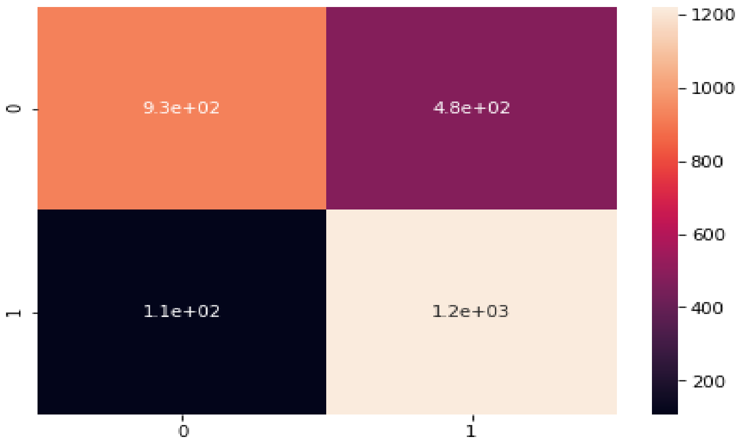


Figure 4.11: Confusion matrix for happy emotion.

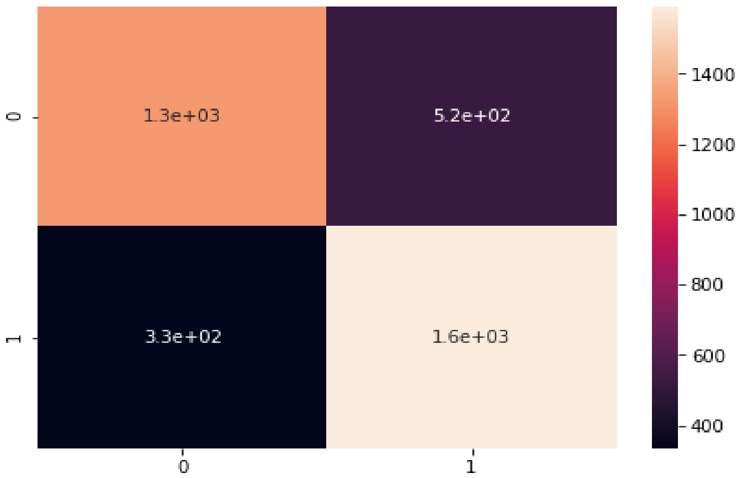


Figure 4.12: Confusion matrix for sad emotion.

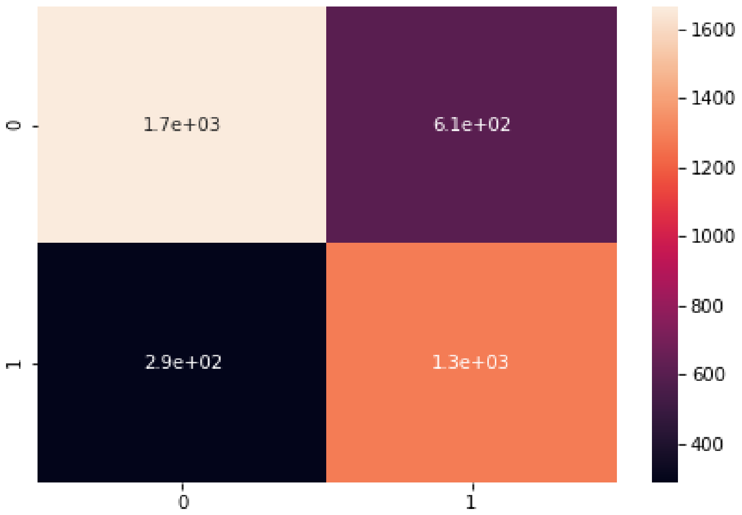


Figure 4.13: Confusion matrix for surprise emotion.

The previous work done by Huynh et. al. had received the following ROC curves when validating their model using the same SASE-FE dataset ((9)):

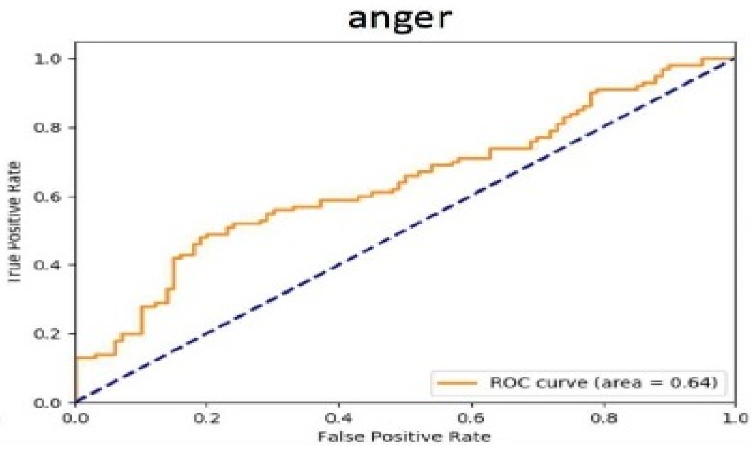


Figure 4.14: ROC curve for anger emotion.

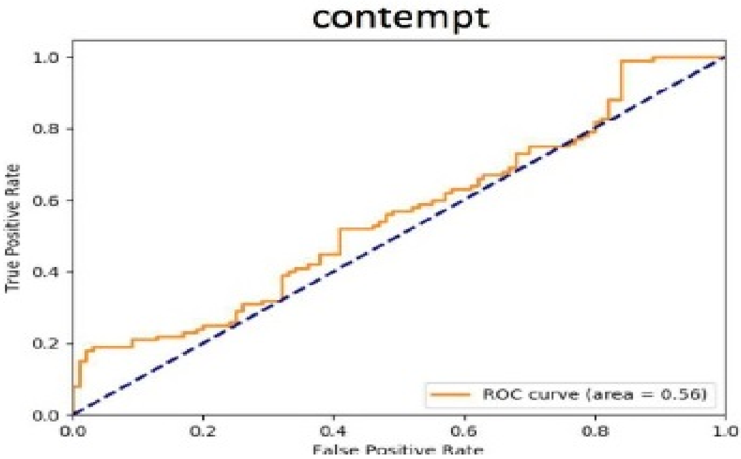


Fig 4.9: ROC curve for anger contempt.

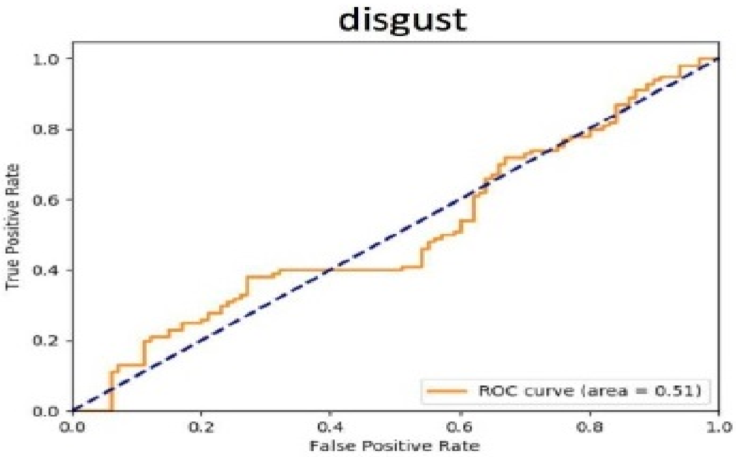


Figure 4.15: ROC curve for disgust emotion.

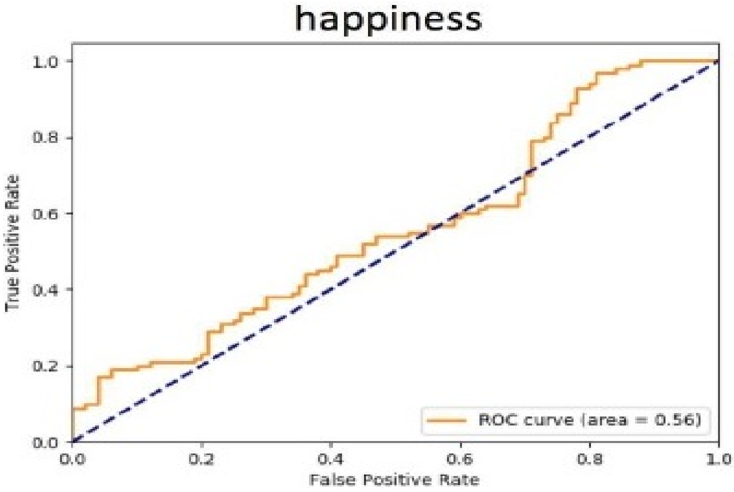


Figure 4.16: ROC curve for happy emotion.

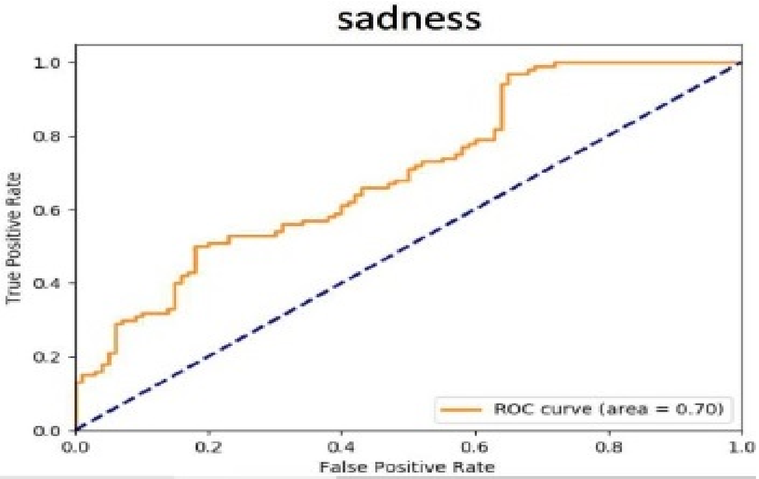


Fig 4.12: ROC curve for sad emotion.

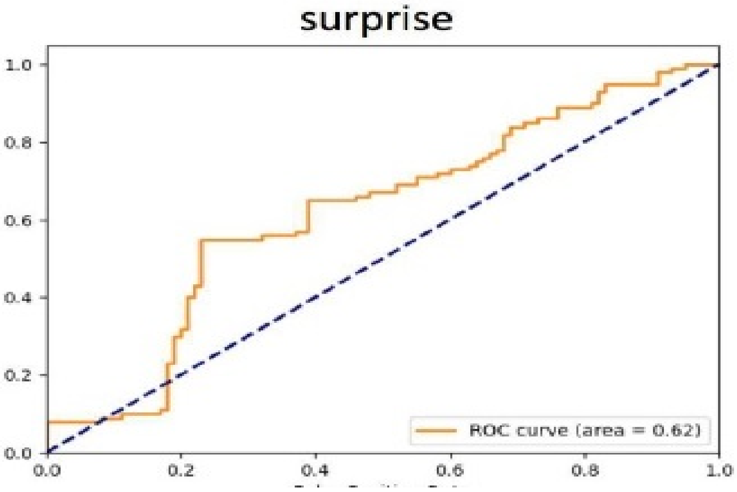


Fig 4.13: ROC curve for surprise emotion.

4.2 Discussion

From the results, we observe that the model performs well when dealing with facial expressions such as anger, happiness, sadness and surprise. The area under the curve was more than 0.80 for each of the emotions, while for contempt and disgust the area under the curve was 0.61 and 0.64 respectively. The accuracy for anger, happiness, sadness and surprise was approximately 75%, for contempt the same being 56% and 61% for disgust. The model achieved highest accuracy for happiness (approximately 79%). Hence, it can be inferred that anger, happiness, sadness and surprise can be easily detected. These findings align with the work of Huynh et. al. where he had made similar observations regarding the feasibility of correctly detecting the above expressions. However, the system failed to satisfactorily detect the genuineness of the emotions of contempt and disgust. I thus infer that the facial expression involved with contempt and disgust vary to a great extent subjectively, whereas the emotions such as anger, surprise, sadness and happiness are quite similar on an average. For example, anger involves frowning, surprise involves wide opened jaws and stretched eyebrows, happiness involves a wide smile with a contraction in the orbicular occuli muscle and sadness involved tightening of the lips and inner brow raise. But individual differences were noticed regarding the portrayal of emotions of contempt and disgust. Another possible reason for the system for not performing well for these two sets of emotions might be due to lack of data, but trying out techniques such as virtual data augmentation, or getting more related data might help in improving the performance of the system.

The usage of Siamese network made it possible to simultaneously compute the features from the set of both true and false images extracted from the videos. By working on both fake set of images of a particular emotion as well as the true set of images for an expression simultaneously the computational time for getting the features of the images was considerably reduced but it is computationally expensive as it uses twice the number of parameters compared to a single ResNet architecture.

Due to shortage of time and limitation on the usage of GPU, k-fold stratified cross validation and techniques such as face alignment or techniques such as 3D face modelling could not be applied. But these techniques can be applied, which might help the model to improve on accuracy. The usage of ResNet architecture has definitely helped in getting improved features from the images and deploying deeper ResNet models such as 100 layers might make the model more accurate or using the architecture of Inception network might even work better.

It can also be understood that the fully connected 10-layers deep neural network has worked better in classifying the images as per genuineness of facial expressions than extreme gradient boosting method used by Huynh et. al.

4.2.1 Limitations of the current research

In this research, the network is only capable of detecting the genuineness of an emotion, but fails to categorise the emotion, i.e., the system is not able to identify which specific emotion is it working the comparison on. To illustrate, for a given a sequence of anger-related images, the model is only capable of finding the ingenuity of a human expression but fails to identify the emotion it is dealing with. Moreover, the model is based on frontal facial features and does not take into account the features from the two facial sides. Including the peripheral face might have result into better accuracy. Next, the model only deals with 6 basic human emotions, whereas recently the list has been expanded to 10 human emotions instead of the previous 7, that are not taken into account (example fear, excitement). Finally, it is a well-accepted fact that along with facial expression of emotions, there are certain bodily non-verbal cues, inclusion of which can be highly beneficial to future researches of this sort (13, 41).

4.2.2 Future Work

In future I would like to expand my model by using a deeper architecture, i.e., by using a deeper ResNet architecture having 100 layers. It would also be very interesting to see how well does an inception layer can perform for this objective. Furthermore, in addition to frontal face features used in this research, side facial features as well as the body movements should also be considered as it forms an important factor for classifying an expression as true or false. I would like to use different Haar cascade features for instance of eyes, mouth, frontal face and body. Compute each features separately and come up with an ensemble approach where the results will be based on the average score from all the individual features, which might be better in predicting the genuineness of a posed facial expression. Finally, the model fails to recognise the emotion it is dealing with. It is desirable that the model can recognise the emotion it is dealing with as well as try and predict the genuineness of the facial expression.

5 Conclusion

Human beings are social animals and communication is a basic necessity of human life. Facial expressions play an active role for effective communicating, while also helping the receiver understand the emotion of the communicator as a whole. This research was dedicated to finding out if a Siamese Neural Network architecture can classify genuine and fake emotions correctly. The SASE-FE dataset was used for this research. From each video of the 50 subjects the frames were extracted and then cropped, and the facial features of these images were collected by a Residual Network (ResNet-50) having 50 layers. A 10 layer deep neural network was used to classify whether the features obtained from the images belong to a true expression of emotion class or a false expression of emotion class. The results achieved by Huynh et. al. served the purpose of the baseline for this study. It was observed that Residual Network architecture, when used in Siamese Neural Network type architecture, performed well, having more that 70 percent of the area under the curve. The results were the best for sadness, happiness and anger, where for each, the area under the curve was more than 80 percent, while for surprise it was 70 percent. On the other hand, the lack of satisfactory results for contempt and disgust could be due to absence of sufficient data and the varied individualistic expression of these two emotions. The greatest difficulty in capturing the exact facial expression was encountered as a few subjects had moved their heads to a great extent, which can be rectified with the help of 3d modelling technique or working on facial alignment. The labour of running the codes could also have been lessened with the availability of more GPU for the run time. My major learning from this research is learning how a Siamese Neural Network, specifically a Residual architectures, learn various features from a given set of images. I have also learnt that few expressions like disgust and contempt are not as universal as compared to surprise, happiness and sadness and therefore are more complex to assess. This system, after needful improvements, can be beneficial to judiciary for forensic and criminal handling, as well as psychological therapeutic processes.

Bibliography

[1] P Rondot. G. b. a. duchenne de boulogne (1806-1875). *Journal of neurology*, 252(7): 866—867, July 2005. ISSN 0340-5354. doi: 10.1007/s00415-005-0874-0. URL `https://doi.org/10.1007/s00415-005-0874-0`.

[2] Charles Darwin. The expression of the emotions in man and animals, new york: D. *Appleton and Company*, 1872.

[3] Rosenberg Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[4] William E Rinn. The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. *Psychological bulletin*, 95(1):52, 1984.

[5] Marilyn L Hill and Kenneth D Craig. Detecting deception in pain expressions: the structure of genuine and deceptive facial displays. *Pain*, 98(1-2):135–144, 2002.

[6] Marian Stewart Bartlett, Gwen C Littlewort, Mark G Frank, and Kang Lee. Automatic decoding of facial movements reveals deceptive pain expressions. *Current Biology*, 24 (7):738–743, 2014.

[7] Erno Mäkinen. Computer vision techniques and applications in human-computer interaction. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, page 354, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581139950. doi: 10.1145/1027933.1028011. URL `https://doi.org/10.1145/1027933.1028011`.

[8] Kaustubh Kulkarni, Ciprian Corneanu, Ikechukwu Ofodile, Sergio Escalera, Xavier Baro, Sylwia Hyniewska, Juri Allik, and Gholamreza Anbarjafari. Automatic recognition of facial displays of unfelt emotions. *IEEE transactions on affective computing*, 2018.

[9] Xuan-Phung Huynh and Yong-Guk Kim. Discrimination between genuine versus fake emotion using long-short term memory with parametric bias and facial landmarks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3065–3072, 2017.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[11] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[12] Guillaume-Benjamin Duchenne and G-B Duchenne de Boulogne. *The mechanism of human facial expression*. Cambridge university press, 1990.

[13] Paul Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1):205–221, 2003.

[14] Paul Ekman, Richard J Davidson, and Wallace V Friesen. The duchenne smile: emotional expression and brain physiology: Ii. *Journal of personality and social psychology*, 58(2):342, 1990.

[15] Michael Lewis, Jeannette M Haviland-Jones, and Lisa Feldman Barrett. *Handbook of emotions*. Guilford Press, 2010.

[16] Eeva Anita Elliott and Arthur M Jacobs. Facial expressions, emotions, and sign languages. *Frontiers in psychology*, 4:115, 2013.

[17] Paul Ekman and Dacher Keltner. Universal facial expressions of emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, pages 27–46, 1997.

[18] Otto Klineberg. Emotional expression in chinese literature. *The Journal of Abnormal and Social Psychology*, 33(4):517, 1938.

[19] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

[20] David Matsumoto, Dacher Keltner, Michelle N Shiota, Maureen O'Sullivan, and Mark Frank. Facial expressions of emotion. 2008.

[21] TL McLellan, JC Wilcke, Lucy Johnston, Richard Watts, and LK Miles. Sensitivity to posed and genuine displays of happiness and sadness: A fmri study. *Neuroscience letters*, 531(2):149–154, 2012.

[22] K Luan Phan, Tor Wager, Stephan F Taylor, and Israel Liberzon. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in pet and fmri. *Neuroimage*, 16(2):331–348, 2002.

[23] Marilyn L Kesler, Anders H Andersen, Charles D Smith, Malcolm J Avison, C Ervin Davis, Richard J Kryscio, Lee X Blonder, et al. Neural substrates of facial emotion processing using fmri. *Cognitive Brain Research*, 11(2):213–226, 2001.

[24] Mary L Phillips, Wayne C Drevets, Scott L Rauch, and Richard Lane. Neurobiology of emotion perception i: The neural basis of normal emotion perception. *Biological psychiatry*, 54(5):504–514, 2003.

[25] Jennifer C Britton, K Luan Phan, Stephan F Taylor, Robert C Welsh, Kent C Berridge, and Israel Liberzon. Neural correlates of social and nonsocial emotions: An fmri study. *Neuroimage*, 31(1):397–409, 2006.

[26] Tatia MC Lee, Ho-Ling Liu, Rumjahn Hoosain, Wan-Ting Liao, Chien-Te Wu, Kenneth SL Yuen, Chetwyn CH Chan, Peter T Fox, and Jia-Hong Gao. Gender differences in neural correlates of recognition of happy and sad faces in humans assessed by functional magnetic resonance imaging. *Neuroscience letters*, 333(1): 13–16, 2002.

[27] Guillaume-Benjamin Duchenne. *Mécanisme de la physionomie humaine: où, Analyse électro-physiologique de l'expression des passions*. J.-B. Baillière, 1876.

[28] Stephen Porter and Leanne Ten Brinke. Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychological science*, 19(5):508–514, 2008.

[29] Paul Ekman and Wallace V Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

[30] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

[31] David Matsumoto and Hyi Sung Hwang. Reading facial expressions of emotion. *Psychological Science Agenda*, 25(5), 2011.

[32] Paul Ekman and Wallace V Friesen. Unmasking the face englewood cliffs. *Spectrum-Prentice Hall, New Jersey*, 1975.

[33] Ross Buck. Prime theory: An integrated view of motivation and emotion. *Psychological review*, 92(3):389, 1985.

[34] Robert H Frank. *Passions within reason: The strategic role of the emotions.* WW Norton & Co, 1988.

[35] Leda Cosmides and John Tooby. Evolutionary psychology and the emotions. *Handbook of emotions*, 2(2):91–115, 2000.

[36] Marc Mehu, Marcello Mortillaro, Tanja Bänziger, and Klaus R Scherer. Reliable facial muscle activation enhances recognizability and credibility of emotional expression. *Emotion*, 12(4):701, 2012.

[37] Pierre Gosselin, Melanie Perron, and Martin Beaupré. The voluntary control of facial action units in adults. *Emotion*, 10(2):266, 2010.

[38] Jerry Weiss. Ekman, p.(2009) telling lies: Clues to deceit in the marketplace, politics, and marriage. new york: Norton, 2011.

[39] Mark G Frank and Paul Ekman. The ability to detect deceit generalizes across different types of high-stake lies. *Journal of personality and social psychology*, 72(6): 1429, 1997.

[40] Elliott D Ross, Luay Shayya, Amanda Champlain, Marilee Monnot, and Calin I Prodan. Decoding facial blends of emotion: Visual field, attentional and hemispheric biases. *Brain and cognition*, 83(3):252–261, 2013.

[41] Stephen Porter, Leanne Ten Brinke, and Brendan Wallace. Secrets and lies: Involuntary leakage in deceptive facial expressions as a function of emotional intensity. *Journal of Nonverbal Behavior*, 36(1):23–37, 2012.

[42] Ikechukwu Ofodile, Ahmed Helmi, Albert Clapés, Egils Avots, Kerttu Maria Peensoo, Sandhra-Mirella Valdma, Andreas Valdmann, Heli Valtna-Lukner, Sergey Omelkov, Sergio Escalera, et al. Action recognition using single-pixel time-of-flight detection. *Entropy*, 21(4):414, 2019.

[43] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[44] Donald Olding Hebb. *The organization of behavior: a neuropsychological theory*. J. Wiley; Chapman & Hall, 1949.

[45] James L McClelland, David E Rumelhart, PDP Research Group, et al. Parallel distributed processing. *Explorations in the Microstructure of Cognition*, 2:216–271, 1986.

[46] Arthur E Bryson. A gradient method for optimizing multi-stage allocation processes. In *Proc. Harvard Univ. Symposium on digital computers and their applications*, volume 72, 1961.

[47] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[48] James L McClelland and David E Rumelhart. *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. MIT press, 1989.

[49] AE Bryson and Yu-Chi Ho. Applied optimal control. 1969. *Blaisdell, Waltham, Mass*, 8(72):14, 1969.

[50] Shunichi Amari. A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*, (3):299–307, 1967.

[51] Paul John Werbos. *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*, volume 1. John Wiley & Sons, 1994.

[52] David Daniel Cox and Thomas Dean. Neural networks and neuroscience-inspired computer vision. *Current Biology*, 24(18):R921–R929, 2014.

[53] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[54] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1): 106, 1962.

[55] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

[56] David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959.

[57] David H Hubel and Torsten N Wiesel. *Brain and visual perception: the story of a 25-year collaboration*. Oxford University Press, 2004.

[58] Kunihiko Fukushima. Neocognitron. *Scholarpedia*, 2(1):1717, 2007.

[59] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

[60] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3642–3649. IEEE, 2012.

[61] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

[62] Les E Atlas, Toshiteru Homma, and Robert J Marks II. An artificial neural network for spatio-temporal bipolar patterns: Application to phoneme classification. In *Neural Information Processing Systems*, pages 31–40, 1988.

[63] John J Weng, Narendra Ahuja, and Thomas S Huang. Learning recognition and segmentation of 3-d objects from 2-d images. In *1993 (4th) International Conference on Computer Vision*, pages 121–128. IEEE, 1993.

[64] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[65] Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339, 1989.

[66] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[67] John B Hampshire II and Alex Waibel. Connectionist architectures for multi-speaker phoneme recognition. In *Advances in neural information processing systems*, pages 203–210, 1990.

[68] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224. IEEE, 2017.

[69] Kouichi Yamaguchi, Kenji Sakamoto, Toshio Akabane, and Yoshiji Fujimoto. A neural network for speaker-independent isolated word recognition. In *First International Conference on Spoken Language Processing*, 1990.

[70] John S Denker, WR Gardner, Hans Peter Graf, Donnie Henderson, Richard E Howard, W Hubbard, Lawrence D Jackel, Henry S Baird, and Isabelle Guyon. Neural network recognizer for hand-written zip code digits. In *Advances in neural information processing systems*, pages 323–331, 1989.

[71] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998.

[72] Wei Zhang et al. Shift-invariant pattern recognition neural network and its optical architecture. In *Proceedings of annual conference of the Japan Society of Applied Physics*, 1988.

[73] Wei Zhang, Kazuyoshi Itoh, Jun Tanida, and Yoshiki Ichioka. Parallel distributed processing model with local space-invariant interconnections and its optical architecture. *Applied optics*, 29(32):4790–4797, 1990.

[74] Wei Zhang, A Hasegawa, K Itoh, and Y Ichioka. Error back propagation with minimum-entropy weights: a technique for better generalization of 2-d shift-invariant nns. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*, volume 1, pages 645–648. IEEE, 1991.

[75] Wei Zhang, Akira Hasegawa, Kazuyoshi Itoh, and Yoshiki Ichioka. Image processing of human corneal endothelium based on a learning network. *Applied Optics*, 30(29): 4211–4217, 1991.

[76] Wei Zhang, Kunio Doi, Maryellen L Giger, Yuzheng Wu, Robert M Nishikawa, and Robert A Schmidt. Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network. *Medical physics*, 21(4): 517–524, 1994.

[77] Sven Behnke. *Hierarchical neural networks for image interpretation*, volume 2766. Springer, 2003.

[78] Kyoung-Su Oh and Keechul Jung. Gpu implementation of neural networks. *Pattern Recognition*, 37(6):1311–1314, 2004.

[79] Dave Steinkraus, Ian Buck, and PY Simard. Using gpus for machine learning algorithms. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 1115–1120. IEEE, 2005.

[80] Kumar Chellapilla, Sidd Puri, and Patrice Simard. High performance convolutional neural networks for document processing. 2006.

[81] Dan Claudiu Cireșan, Ueli Meier, Luca Maria Gambardella, and Jürgen Schmidhuber. Deep, big, simple neural nets for handwritten digit recognition. *Neural computation*, 22(12):3207–3220, 2010.

[82] Dan Claudiu Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[83] Dan Cireșan, Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber. A committee of neural networks for traffic sign classification. In *The 2011 international joint conference on neural networks*, pages 1918–1921. IEEE, 2011.

[84] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1): 98–113, 1997.

[85] Masakazu Matsugu, Katsuhiko Mori, Yusuke Mitari, and Yuji Kaneda. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks*, 16(5-6):555–559, 2003.

[86] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In *International workshop on human behavior understanding*, pages 29–39. Springer, 2011.

[87] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[88] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[89] Thomas Huang. Computer vision: Evolution and promise. 1996.

[90] Lawrence G Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963.

[91] John Aloimonos. Purposive and qualitative active vision. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume 1, pages 346–360. IEEE, 1990.

[92] Patricia Kitcher. Marr's computational theory of vision. *Philosophy of Science*, 55(1): 1–24, 1988.

[93] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[94] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[95] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.

[96] Martin A Fischler and Robert A Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on computers*, 100(1):67–92, 1973.

[97] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3025–3032, 2013.

[98] Xudong Cao, David Wipf, Fang Wen, Genquan Duan, and Jian Sun. A practical transfer learning algorithm for face verification. In *Proceedings of the IEEE international conference on computer vision*, pages 3208–3215, 2013.

[99] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, UMass Amherst technical report, 2010.

[100] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE, 2012.

[101] Junjie Yan, Xuzong Zhang, Zhen Lei, and Stan Z Li. Face detection by structural models. *Image and Vision Computing*, 32(10):790–799, 2014.

[102] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *European conference on computer vision*, pages 109–122. Springer, 2014.

[103] Chang Huang, Haizhou Ai, Yuan Li, and Shihong Lao. High-performance rotation invariant multiview face detection. *IEEE Transactions on pattern analysis and machine intelligence*, 29(4):671–686, 2007.

[104] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. A deep pyramid deformable part model for face detection. In *2015 IEEE 7th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–8. IEEE, 2015.

[105] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z Li. Convolutional channel features. In *Proceedings of the IEEE international conference on computer vision*, pages 82–90, 2015.

[106] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z Li. Aggregate channel features for multi-view face detection. In *IEEE international joint conference on biometrics*, pages 1–8. IEEE, 2014.

[107] Sachin Sudhakar Farfade, Mohammad J Saberian, and Li-Jia Li. Multi-view face detection using deep convolutional neural networks. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 643–650, 2015.

[108] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5325–5334, 2015.

[109] Niall O'Mahony, Trevor Murphy, Krishna Panduru, Daniel Riordan, and Joseph Walsh. Adaptive process control and sensor fusion for process analytical technology. In *2016 27th Irish Signals and Systems Conference (ISSC)*, pages 1–6. IEEE, 2016.

[110] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

[111] Régis Vaillant, Christophe Monrocq, and Yann Le Cun. Original approach for the localisation of objects in images. *IEE Proceedings-Vision, Image and Signal Processing*, 141(4):245–250, 1994.

[112] Henry A Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Transactions on pattern analysis and machine intelligence*, 20(1): 23–38, 1998.

[113] Mengyi Liu, Shaoxin Li, Shiguang Shan, Ruiping Wang, and Xilin Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Asian conference on computer vision*, pages 143–157. Springer, 2014.

[114] Gary B Huang, Honglak Lee, and Erik Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2518–2525. IEEE, 2012.

[115] Pablo Barros, Cornelius Weber, and Stefan Wermter. Emotional expression recognition with a cross-channel convolutional neural network for human-robot

interaction. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 582–587. IEEE, 2015.

[116] Gregor Domes, Ekkehardt Kumbier, Markus Heinrichs, and Sabine C Herpertz. Oxytocin promotes facial emotion recognition and amygdala reactivity in adults with asperger syndrome. *Neuropsychopharmacology*, 39(3):698–706, 2014.

[117] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[118] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2), 2016.

[119] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991, 2015.

[120] Xi Ouyang, Shigenori Kawaai, Ester Gue Hua Goh, Shengmei Shen, Wan Ding, Huaiping Ming, and Dong-Yan Huang. Audio-visual emotion recognition using deep transfer learning and multiple temporal models. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 577–582, 2017.

[121] Duc Minh Vo, Akihiro Sugimoto, and Thai Hoang Le. Facial expression recognition by re-ranking with global and local generic features. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 4118–4123. IEEE, 2016.

[122] Shulin Xu, Nan Pu, Li Qian, and Guoqiang Xiao. Combination of pyramid cnn representation and spatial-temporal representation for facial expression recognition. In *CCF Chinese Conference on Computer Vision*, pages 40–50. Springer, 2017.

[123] Behzad Hasani and Mohammad H Mahoor. Facial expression recognition using enhanced deep 3d convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 30–40, 2017.

[124] Zhenghao Li, Song Wu, and Guoqiang Xiao. Facial expression recognition by multi-scale cnn with regularized center loss. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3384–3389. IEEE, 2018.

[125] Jin Zhang, Xue Mei, Huan Liu, Shenqiang Yuan, and Tiancheng Qian. Detecting negative emotional stress based on facial expression in real time. In *2019 IEEE 4th*

International Conference on Signal and Image Processing (ICSIP), pages 430–434. IEEE, 2019.

[126] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[127] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[128] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *Proceedings. international conference on image processing*, volume 1, pages I–I. IEEE, 2002.

[129] Gary Bradski. The opencv library. *Dr Dobb's J. Software Tools*, 25:120–125, 2000.

[130] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[131] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[132] Kaiming He and Jian Sun. Convolutional neural networks at constrained time cost. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5353–5360, 2015.

[133] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[134] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[135] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[136] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[137] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008.

[138] Dong Chen, Xudong Cao, Liwei Wang, Fang Wen, and Jian Sun. Bayesian face revisited: A joint formulation. In *European conference on computer vision*, pages 566–579. Springer, 2012.

[139] Thomas Berg and Peter N Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. In *Bmvc*, volume 2, page 7. Citeseer, 2012.

[140] Zara Ambadar, Jonathan W Schooler, and Jeffrey F Cohn. Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. *Psychological science*, 16(5):403–410, 2005.

[141] Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

[142] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133, 2014.

[143] Paul Ekman. Facial action coding system. 1977.

[144] Jeffrey F Cohn and Karen Schmidt. The timing of facial motion in posed and spontaneous smiles. In *Active Media Technology*, pages 57–69. World Scientific, 2003.

[145] Patrick Lucey, Jeffrey F Cohn, Iain Matthews, Simon Lucey, Sridha Sridharan, Jessica Howlett, and Kenneth M Prkachin. Automatically detecting pain in video through facial action units. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(3):664–674, 2010.

[146] Michel F Valstar, Hatice Gunes, and Maja Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 38–45, 2007.

[147] Wataru Sato, Takanori Kochiyama, Sakiko Yoshikawa, Eiichi Naito, and Michikazu Matsumura. Enhanced neural activity in response to dynamic facial expressions of emotion: an fmri study. *Cognitive Brain Research*, 20(1):81–91, 2004.

[148] Rachael E Jack and Philippe G Schyns. The human face as a dynamic tool for social communication. *Current Biology*, 25(14):R621–R634, 2015.

[149] Stephen V Stehman. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1):77–89, 1997.

[150] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[151] Jerome Fan, Suneel Upadhye, and Andrew Worster. Understanding receiver operating characteristic (roc) curves. *Canadian Journal of Emergency Medicine*, 8(1):19–20, 2006.

[152] Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

[153] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2544–2550. IEEE, 2010.