An ANN Based Combined Classifier Approach for Facial Emotion Recognition

by

EKİN YAĞIŞ

Submitted to
the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

SABANCI UNIVERSITY

July 2018

AN ANN BASED COMBINED CLASSIFIER APPROACH FOR FACIAL EMOTION RECOGNITION

APPROVED BY:

Prof. Dr. Mustafa Ünel.. Mustafa Ünel

(Thesis Supervisor)

Asst. Prof. Dr. Hüseyin Özkan...

Assoc. Prof. Dr. Şeref Naci Engin..

DATE OF APPROVAL: 30/07/2018

ABSTRACT

AN ANN BASED COMBINED CLASSIFIER APPROACH FOR FACIAL EMOTION RECOGNITION

EKIN YAGIS

Mechatronics Engineering M.Sc. Thesis, August 2018

Thesis Supervisor: Prof. Dr. Mustafa Ünel

Keywords: Emotion Recognition, Facial Expression Analysis, Facial Action Coding System, Classification, Artificial Neural Networks (ANN), Logarithmic Opinion Pool (LOP)

Facial expressions are the simplest reflections of human emotions, which are at the same time an integral part of any communication. Over the last decade, facial emotion recognition has attracted a great deal of research interest due to its various applications in the fields such as human computer interaction, robotics and data analytics.

In this thesis, we present a facial emotion recognition approach that is based on facial expressions to classify seven emotional states: neutral, joy, sadness, surprise, anger, fear and disgust. To perform classification, two different facial features called Action Units (AUs) and Feature Point Positions (FPPs) are extracted from image sequences. A depth camera is used to capture image sequences collected from 13 volunteers to classify seven emotional states. Having extracted two sets of features, separate artificial neural network classifiers are trained. Logarithmic Opinion Pool (LOP) is then employed to combine the decision probabilities coming from each classifier. Experimental results are quite promising and establish a basis for future work on the topic.

ÖZET

YAPAY SİNİR AĞLARI TEMELLİ BİRLEŞİK SINIFLANDIRICILAR İLE YÜZ İFADELERİNDEN DUYGU TANIMA

EKİN YAĞIŞ

Mekatronik Mühendisliği Yüksek Lisans Tezi, Ağustos 2018

Tez Danışmanı: Prof. Dr. Mustafa Ünel

Anahtar Kelimeler: Duygu Tanıma, Yüz İfade Analizi, Yüz Hareketleri Kodlama Sistemi, Sınıflandırma, Yapay Sinir Ağları, Logaritmik Düşünce Havuzu

Yüz ifadeleri herhangi bir tür iletişimin temeli olmakla beraber insan duygularının en basit yansımasıdır. İnsan-makine etkileşiminden, robotiğe ve veri analitiğine kadar pek çok uygulama alanı olması nedeniyle duygu durumu sınıflandırılması son on yılda çokça araştırılmıştır.

Bu tez çalışmasında yüz ifadesi temelli yedi basit duygu durumunun (ifadesizlik, neşe, mutsuzluk, sürpriz, kızgınlık, korku ve iğrenme) sınıflandırılması için yeni bir yaklaşım geliştirilmiştir. Bu amaçla her bir ifade için alınan seri görüntülerden 'hareket birimleri' ve 'nokta pozisyonları' denilen iki farklı öznitelik çıkartılmıştır. Onüç farkı gönüllüden elde edilen yüz ifadelerinin kayıt edilmesinde bir derinlik kamerası kullanılmıştır. Özniteliklerin çıkartılmasının ardından duygu durumlarının sınıflandırılması için iki ayrı yapay sinir ağı eğitilmiştir. Sonrasında logaritmik düşünce havuzu adı verilen bir olasılıksal modelleme ile her bir sınıflandırıcıdan gelen karar olasılıkları birleştirilmiştir. Sınıflandırma sonuçları umut verici olup gelecekte bu konuda yapılacak çalışmalar için bir baz oluşturmaktadır.

$\ll$ *Aileme* $\gg$

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis advisor, Prof.Dr. Mustafa Unel for his invaluable academic guidance, help and support. He inspired me in pursuing the Machine Learning track and introduced me to Computer Vision. I am greatly indebted to him for his unique supervision throughout my Master study at Sabancı University.

I would gratefully thank Asst. Prof. Dr. Hüseyin Özkan and Assoc. Prof. Dr. Şeref Naci Engin for their interest in my work and spending their valuable time to serve as my jurors.

I would also like to thank every single member of Control, Vision and Robotics Laboratory for being good companies during my graduate study. Many thanks to Gökhan Alcan, Diyar Khalis Bilal, Sanem Evren and Zain Fuad for their help, friendship and sharing all they know; to Hande Karamahmutoğlu and Yağmur Yıldızhan for the wonderful times we shared and always being there for me and all mechatronics laboratory members I wish I had the space to acknowledge in person for their great friendship throughout my Master study.

I would like to thank my family, especially Mom and Dad, for invaluable love, caring and continuous support they have given me from the beginning of my life. I could not have done it without them.

Finally, I would like to thank Arın for all the love, comfort, and guidance he has given me throughout my time in graduate school.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The human face is a complex structure that allows diverse and subtle facial expressions. During conversations, the first thing that brings to the others' attention is our faces [21]. Together with gestures, speech and physiological parameters, facial expressions may reveal a lot of information about people's feelings. Therefore, facial expressions are essential part of human communication.

In 1968, Mehrabian concluded in his research that 55% of message conveying information about feelings and attitudes is transmitted through facial expressions [21]. Reading emotion cues from facial expressions is one of the cognitive functions that our brain performs quite accurately and efficiently. However, reading emotional reactions in humans is not a trivial task for machines. There are many challenges that may come from the high variability of data. Still, it is vital for machines to get these human-oriented skills by capturing and understanding expressions of emotion such as joy, anxiety, and sorrow if they are to be an indispensable part of human life. Thus, designing an automated facial emotion recognition (FER) system is pivotal to the development of more effective human-computer interaction systems in artificial intelligence (AI).

Over the last decades, research on facial emotion has been gaining a lot of attention with applications not only in the psychology and computer sciences, but also in market research. It can be foreseen that emotionally aware AI systems will gain a competitive advantage over AI with only computational capabilities, especially when it comes to customer experience.

For instance, in the automotive industry, it is expected that driverless cars will be the future and make roads safer in time. To integrate these autonomous systems into our lives and increase the passenger's comfort, his/her emotional state can be inferred using facial emotion recognition. In case of anger, a virtual assistant may motivate the passenger to take a deep breath, play the driver's preferred playlist or suggest a stop along the way. Moreover, the vehicle can change its environmental conditions such as lighting and heating considering passenger's mode and drowsiness.

On the other hand, facial emotion recognition can also be used to capture struggling students in online tutoring sessions. The system can be trained to recognize whether students are engaged in content or not using their facial expressions, so that the topics in which the students are having trouble can be differentiated. To enrich the online learning experience, we can take the advantages of this kind of systems.

For marketers, being able to understand how a customer feels is essential. Instead of gathering feedback from questionnaires and surveys, marketers will soon be able to use the advantage of facial emotion recognition technology. FER provides an opportunity for companies to gain a greater insight about customers and their interests/needs causing an extreme personalization on advertising. Estimating the internal state of shoppers can help to make better business decisions by improving product and service offerings [22].

1.1 Contributions of the Thesis

Emotion recognition problems in AI generally have two components: sensing and adapting. This thesis aims to design an emotion recognition system which focuses on sensing part using a depth camera. We propose a novel method for facial expression-based emotion recognition using action units (AUs) and 3D feature point positions (FPPs). Kinect v2 is used to capture image sequences collected from 13 volunteers to classify seven emotional states. For each frame, 1347 3D facial points and 17 action units are acquired using Face Tracking SDK [23]. Key facial points are selected to reduce the computational cost. Finally, two different neural network classifier classifiers are trained where the inputs of the classifiers are AUs and FPPs respectively. Outputs of individual classifiers are then combined by a decision level fusion algorithm in a probabilistic manner using Logarithmic Opinion Pool (LOP).

The contributions of this thesis are as follows:

- A database of facial images for emotion recognition purposes is created. To the best of our knowledge, there is little previous research on Kinect-based emotion recognition using both action units and feature point positions as features. Hence, there is no public database for evaluating the performance of such emotion recognition systems. We will make this database available for other researchers to use in their work.

- Two different features, namely Action Units (AUs) and Feature Point Positions (FPPs) are extracted from these images and separate Neural Network classifiers are trained.

- Decision level fusion is performed on the outputs of the neural network classifiers using Logarithmic Opinion Pool (LOP) [24–26].

- The proposed algorithm is tested in several scenarios including subject dependent and independent situations. Experimental results are quantified by constructing confusion matrices.

1.2 Outline of the Thesis

Chapter 2 presents the literature survey and theoretical background of facial expressions and emotions. An overview of facial expression measurement and recognition systems used in literature are also provided. **Chapter 3** is on dataset generation and sensors. Detailed explaination of dataset generation procedure as well as the Kinect sensor are introduced. **Chapter 4** details feature extraction processes and dimensionality reduction algorithms. Furthermore, it describes our proposed classification method. Experimental results are presented in **Chapter 5**. Finally, the thesis is concluded in **Chapter 6** and possible future directions are indicated.

1.3 Publications

- E. Yagis, M.Unel, "Facial Expression Based Emotion Recognition Using Neural Networks", International Conference on Image Analysis and Recognition (ICIAR), Povoa de Varzim, Portugal, June 27-29, 2018, Lecture Notes in Computer Science, Vol. 10882, July 2018.

- E.Yagis, M.Unel, "Kinect Based Facial Emotion Recognition Using Fusion of Facial Point Positions and Action Units", Journal Paper(under preparation)

Chapter 2

Literature Survey and Background

2.1 Facial Expression and Emotions

Investigating human emotions is not a novelty if viewed from the perspective of human history. The first studies conducted on emotions can be traced back to the 17th century. It was revolutionary when Descartes first insisted that there must be a relationship between mental processes and body responses. This was controversial in a sense that during that time no device was available to measure such connection. One of the most influential works in the area is Charles Darwin's "The Expression of Emotions in Man and Animals" book [27]. In that book, Darwin claims that facial expressions of emotions have a certain level of universality and evolutionary meaning for survival. Being inspired by this work, Paul Ekman, Wallace Friesen and Carroll Izard, pioneers in this field, conducted several cross-cultural studies known today as "universality studies", on non-verbal expression of emotions.

In 1972 Ekman [28] and Friesen [29] conducted an experiment on facial expression behavior on the faces of Japanese and American students. The subjects were exposed to stressful films. Throughout the films, variances of their facial expression were measured and noted. After these experiments, Ekman and Friensen found that both American and Japanese students showed the same expression when watching emotion-eliciting movies. The subjects were in the condition where they thought they were alone and unobserved. Moreover, they observed an isolated tribe in New Guinea as well as civilized people and came to the conclusion that six categories of emotions, namely happiness, sadness, anger, disgust, fear and surprise can be considered as universal [30, 31].

2.2 Facial Expression Measurement

2.2.1 Physiology of Facial Expressions

The human face contains 20 flat skeletal muscles which are controlled by a cranial nerve. They are located under the skin, mainly near mouth nose and eyes. Facial muscles are unique and different compared to the other groups of muscles in the body. Unlike the other skeletal muscles, they do not move joints and bones but the skin causing facial surface deformations which can be thought as expression.

2.2.2 Facial Action Coding System (FACS)

In the 70s, connections between one's facial muscle movements and his/her psychological state were seriously questioned. There was a need for the development of a coding schema for measuring and classifying facial emotions. At that time, several systems had been developed to solve that problem. Among all the efforts, the Facial Action Coding System (FACS) developed by Ekman and Friesen [30, 31] and the Maximally Discriminative Facial Movement Coding System (MAX) developed by Izard in 1979 [32] were the most prominent schemas. After the development of these

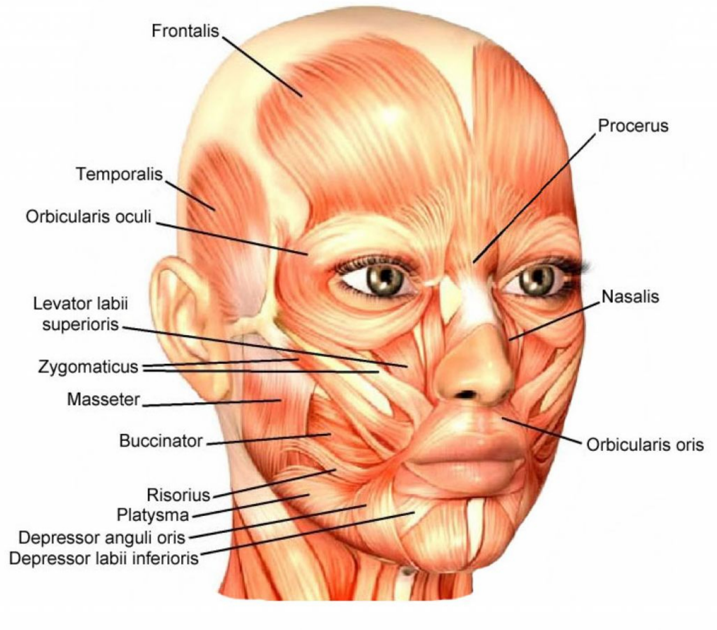


Figure 2.1: Muscles of facial expressions [1])

systems, the research on facial expression and emotion anaysis has attracted a lot of research attention and gained pace.

So far, several methods have been proposed for recognizing and classifying facial emotions. Most of these research works are based on the Facial Action Coding System (FACS) developed by Ekman et al. [30]. The reason behind why FACS was more popular is that rather than focusing on the meaning of emotions it was a comprehensive system based on the anatomical structure of the face. FACS has an enumeration regarding all the muscles in the face responsible for movement whereas MAX is limited to several number of muscle movements which are only related to emotions. Moreoever, FACS scores head and eye movements as well as the muscles.

In 1978, Ekman et al. found out that each emotion results in specific muscle movement [30]. For instance anger affects whole body by activating the chain of reactions in our brain. It increases the heart rate, blood pressure as well as body temperature. The physiological response causes a variety of features which can be observed in face such as wrinkles on the forehead and lifted eyebrows. On the other hand, happiness reveals itself as a smile on the face which is caused by raised cheeks and pulled lip corners. The specific facial features related with each emotion are the following:

1. Joy - Eyes open, cheek raised, lip corners raised, possibly visible teeth, wrinkles outside the eye corners.
2. Sadness - Inner part of eyebrows pulled down, eyes open, lip corners depressed.
3. Surprise - Eyes wide open, jaw dropped, and mouth wide open.
4. Anger - Eyebrows lowered, eyes slightly open, lip corners slightly depressed, tensed jaw.
5. Fear - Eyebrows lowered, mouth open, lips tight and eyes slightly open.
6. Disgust - Eyebrows lowered, eyes almost shut, upper lip lifted, tensed jaw, nose wrinkled.

They developed a method called Facial Action Coding System (FACS) to characterize the physical expression of emotions. FACS is a system which is solely based on anatomical structure of the human face. It characterizes facial activity using the action of a muscle or a group of muscles known as Action Units (AU). For instance, Orbicularis oculi and pars orbitalis muscles are active in the movement of cheeks. FACS consists of 44 AUs of which 12 are for upper face, 18 are for lower face plus another 14 for head or eye movements.

FACS has also a scoring system which is based on the intensity of each facial action, on an A to E scale. An intensity score "A" means that the coder is able to detect slight movement whereas "E" represent the highest movement of specific action unit.

Training human experts to manually score the action units is costly and time consuming. There are numerous studies on automatic recognition of action units for facial expression analysis. The recent advances in machine learning and image processing open up the possibility of extracting action units from facial images. To this end accurate extraction of facial features is a crucial step.

In 1999, Chowdhury et al. [33] tried to identify several basic facial action such as blinking, movements of mouth and eyes using Hidden Markov Models and multidimensional receptive field histograms. However, this model was not able to differentiate relatively complex movement like eyebrow movements. Ohya et al.[34] created a system to recognize head movements. The system could identify major head movements such as shaking and nodding; however, rest of the action units related to facial action was missing. In 2000, Lien et al. [35] detected various action units using dense flow and feature point tracking. In 2001, Tian et al. [12] used multistate templates to detect features like mouth, cheeks, eyebrows, eyes etc. Neural network classifier was then utilized to recognize the facial action units. They achieved to identify sixteen facial actions. In the same year, Cowie et al. [36] introduced a semi-automatic system for identification of action units using Facial Animation Parameter Units. In 2006, Bartlett et al. [2] published a work entitled "Automatic Recognition of Facial Actions in Spontaneous Expressions". In this work, they first detected frontal faces in the video stream and coded each frame with respect to 20 Action units. The approach utilizes support vector machines (SVMs) and AdaBoost and the output of the classifier is the frame by frame action unit intensity.

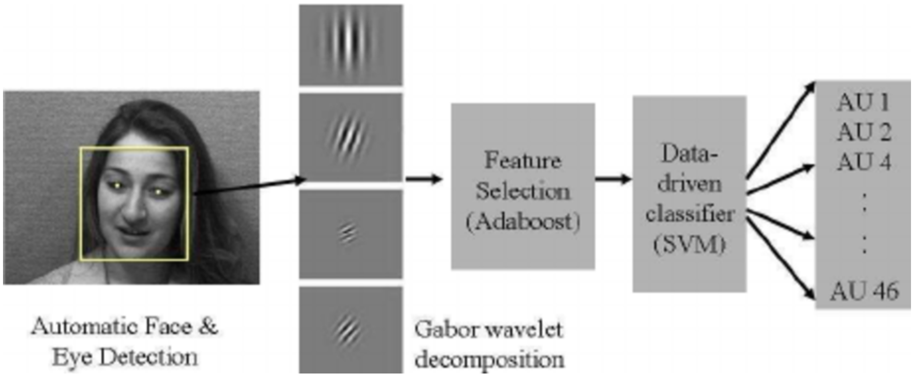


Figure 2.2: Overview of fully automated facial action coding system designed by Bartlett et al. [2])

In the same year, Michel Valstar and Maja Pantic [37] published another work on automatic facial action unit detection and temporal analysis. They first used a facial point localization method which employs GentleBoost templates built from

Gabor wavelet features. After exploiting a particle filtering scheme, SVM classifier was trained on a subset of most informative spatio-temporal features selected by AdaBoost to recognize action units and their temporal segments. They succesfully achieved to classify 15 action units with a mean agreement rate of 90.2% with human FACS coders.

Another influential work in the area is an animation model built from the FACS, for coding human faces called CANDIDE-3. The CANDIDE model was first developed by Mikael Rydfalk at Linköping University in 1987 [3]. It is a parameterised face mask which is controlled by mapping global and local Action Units (AUs) to the alteration of the vertices of the mask. The global action units are the ones that account for the rotations around x, y and z axes whereas the local ones regulate the mimics of the face in order for different expressions to be obtained. The CANDIDE-3 models a human face as a polygon object by using 113 vertices and 168 surfaces. The face mask model can be seen in Figure 2.3 below. The Kinect Face Tracking SDK is also based on the CANDIDE- 3 model. The system is described in detail in Section 3: Dataset Generation and Sensors.

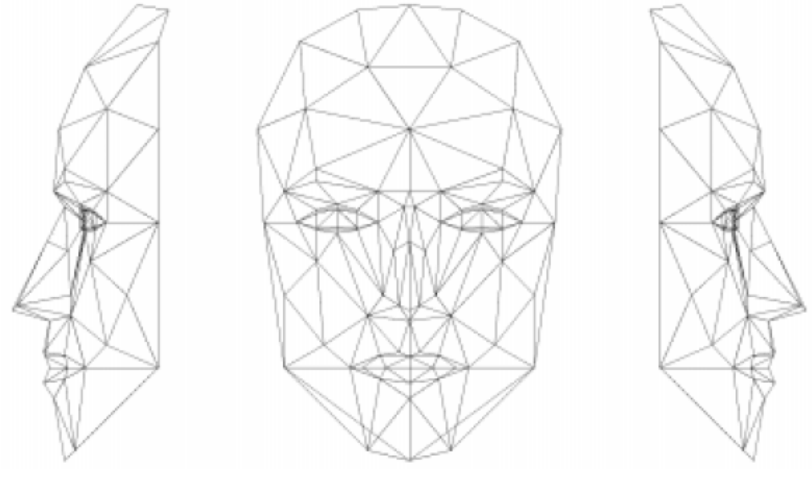
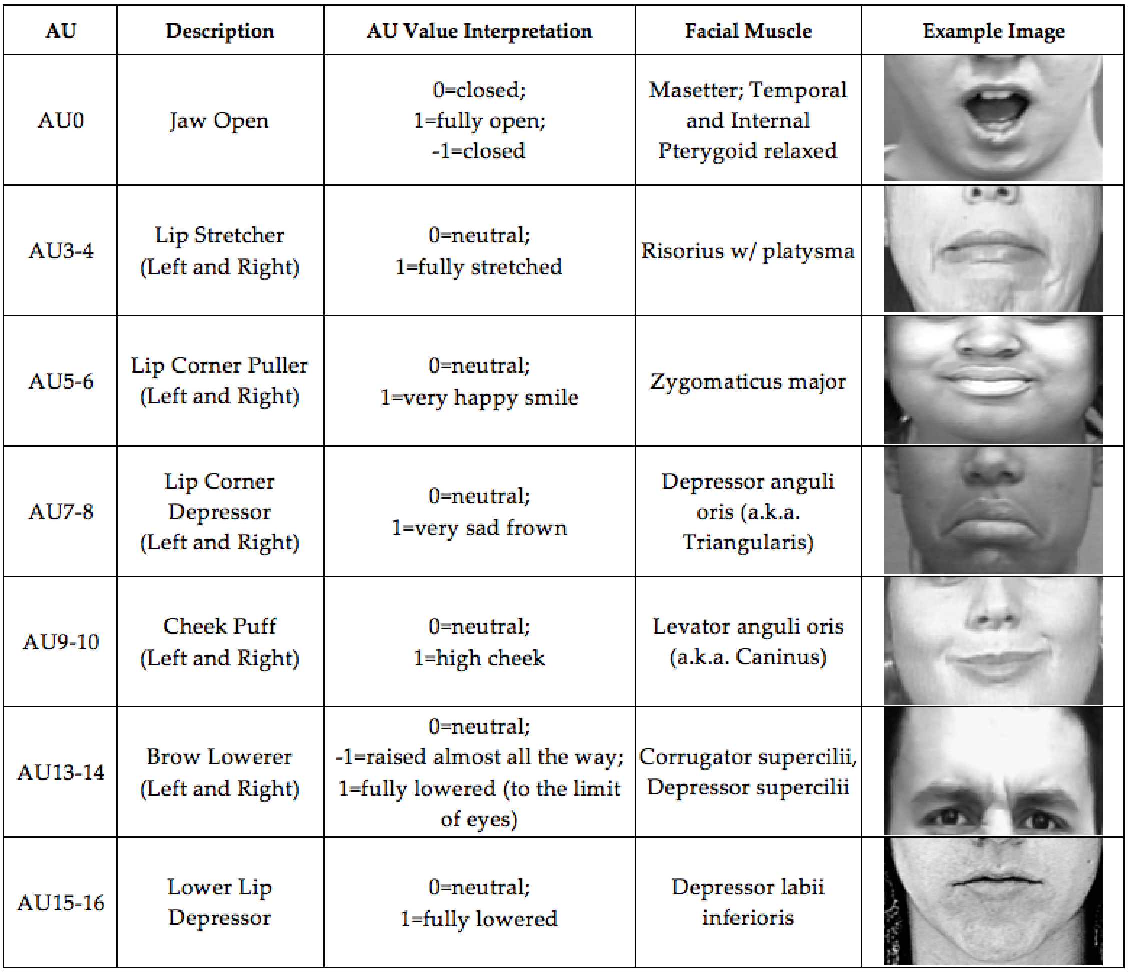


Figure 2.3: CANDIDE-3 face model [3]

Using the earlier version of Kinect sensor only 6 action units could be detected, whereas 17 action units (AUs) can be tracked with the new Kinect v2 and the high definition face tracking API. Out of 17 AUs that are tracked, 13 AUs, their descriptions and the names of the specific facial muscles are visualized in Figure 2.4.



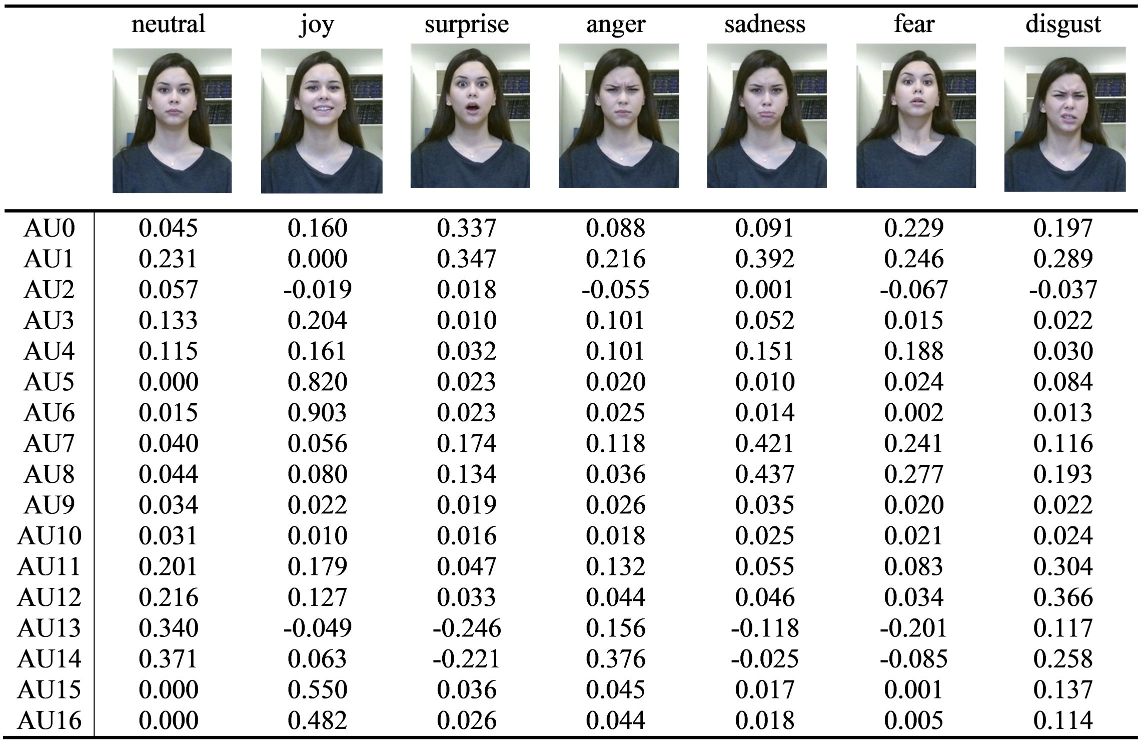
AU	Description	AU Value Interpretation	Facial Muscle	Example Image

AU0	Jaw Open	0=closed;
		1=fully open;
		-1=closed
			Masetter; Temporal and Internal Pterygoid relaxed	

AU3-4	Lip Stretcher (Left and Right)	0=neutral; 1=fully stretched	Risorius w/ platysma	
AU5-6	Lip Corner Puller (Left and Right)	0=neutral; 1=very happy smile	Zygomaticus major	
AU7-8	Lip Corner Depressor (Left and Right)	0=neutral; 1=very sad frown	Depressor anguli oris (a.k.a. Triangularis)	
AU9-10	Cheek Puff (Left and Right)	0=neutral; 1=high cheek	Levator anguli oris (a.k.a. Caninus)	
AU13-14	Brow Lowerer (Left and Right)	0=neutral; -1=raised almost all the way; 1=fully lowered (to the limit of eyes)	Corrugator supercilii, Depressor supercilii	
AU15-16	Lower Lip Depressor	0=neutral; 1=fully lowered	Depressor labii inferioris	

FIGURE 2.4: Several action units together with their description and interpretation (face images are taken from CK+ dataset [4])

14 out of 17 AUs are expressed as a numeric weight varying between 0 and 1 whereas the remaining 3, Jaw Open, Right Eyebrow Lowerer, and Left Eyebrow Lowerer, vary between -1 and +1. For example, if the value of AU 13 is -1 that means left brow is raised fully in most of the cases to express agreement, surprise or fear whereas +1 means that it is lowered to the limit of the eyes showing anger or frustration. Figure 2.5 shows the change of action units for seven emotional states (neutral, joy, surprise, anger, sadness, fear and disgust) of one participant.



	neutral	joy	surprise	anger	sadness	fear	disgust
AU0	0.045	0.160	0.337	0.088	0.091	0.229	0.197
AU1	0.231	0.000	0.347	0.216	0.392	0.246	0.289
AU2	0.057	-0.019	0.018	-0.055	0.001	-0.067	-0.037
AU3	0.133	0.204	0.010	0.101	0.052	0.015	0.022
AU4	0.115	0.161	0.032	0.101	0.151	0.188	0.030
AU5	0.000	0.820	0.023	0.020	0.010	0.024	0.084
AU6	0.015	0.903	0.023	0.025	0.014	0.002	0.013
AU7	0.040	0.056	0.174	0.118	0.421	0.241	0.116
AU8	0.044	0.080	0.134	0.036	0.437	0.277	0.193
AU9	0.034	0.022	0.019	0.026	0.035	0.020	0.022
AU10	0.031	0.010	0.016	0.018	0.025	0.021	0.024
AU11	0.201	0.179	0.047	0.132	0.055	0.083	0.304
AU12	0.216	0.127	0.033	0.044	0.046	0.034	0.366
AU13	0.340	-0.049	-0.246	0.156	-0.118	-0.201	0.117
AU14	0.371	0.063	-0.221	0.376	-0.025	-0.085	0.258
AU15	0.000	0.550	0.036	0.045	0.017	0.001	0.137
AU16	0.000	0.482	0.026	0.044	0.018	0.005	0.114

Figure 2.5: Labeled facial expressions with corresponding action units

2.3 Facial Emotion Recognition Systems

The conventional facial emotion detection system consists of several steps which are shown in Figure 2.6 below. Given input images, first step is to detect face and facial landmarks such as eyes, mouth and nose. Then, a key feature extraction step is employed. Lastly, classification based on several spatial and temporal features is performed using various machine learning algorithms such as support vector machine (SVM), AdaBoost and random forest or neural networks.

2.3.1 Characteristics of an Ideal System

As effortless as it sounds for humans to detect and recognize facial emotions, having systems which can understand emotions are not that easy. Researchers have tried

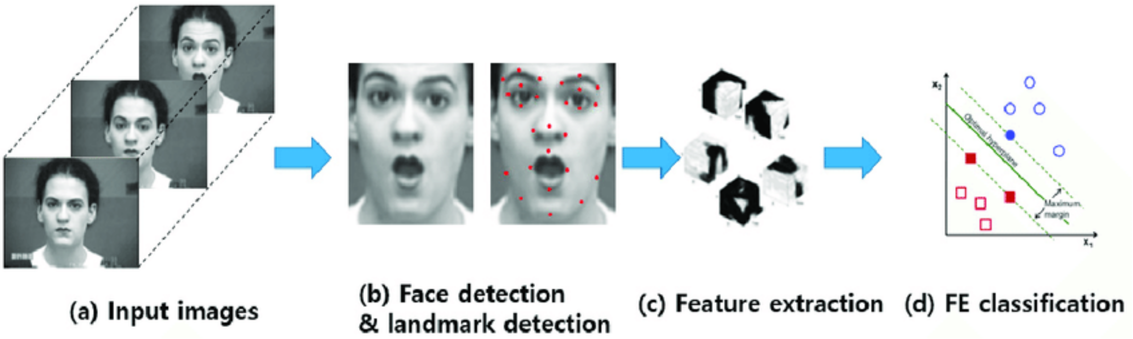


Figure 2.6: Block diagram for conventional FER approaches [5] (face images are taken from CK+ dataset [4])

to solve the way our visual system works in order to come up with a list of some attributes of an ideal automatic FER system. In the book published in 2005, entitled Handbook of Facial Recognition, Tian et al.[38] provided the following properties:

- working in real life scenarios, with any type of images
- being able to recognize both mimicked emotions and genuine human emotions
- being independent of person, gender and age
- being invariable to changes in lighting conditions
- being able to detect and track facial features

2.3.2 Facial Emotion Recognition Approaches

So far, facial emotion recognition systems can be classified as image-based, video-based, and 3D surface-based methods [39]. In image-based approaches, features are usually extracted from the global face region [40] or different face regions containing different types of information [41, 42]. For instance, Happy et al. [40] extracted a local binary pattern (LBP) histogram of different block sizes from a global face region as the feature vectors and classified several facial expressions. However, since different face regions have different levels of importance for emotion recognition, the recognition accuracy tends to be unstable because local variations of the facial parts are not reflected to the feature vector. Ghimire et al.[43] utilized region-specific appearance features by dividing the face region into domain-specific local regions which results in an improvement in the recognition accuracy.

Apart from 2D image-based emotion recognitions, 3D and 4D (dynamic 3D) recordings are increasingly used in FER research. 3D facial expression recognition generally consists of feature extraction and classification. One thing to note is that 3D approaches can also be divided into two categories based on the nature of the data: dynamic and static. In static systems, feature extraction is performed from statistical models such as deformable model, active shape model and distance-based features whereas in dynamic systems 3D motion-based features are extracted from image sequences [5].

Over the past decades, some researchers have used Kinect sensor to recognize emotions. Kinect is a high speed optical sensor with the abilities of both traditional RGB cameras and 3D scanning equipment. It is affordable for many applications, fast in scanning, and compact in size. Mostly, it can be said that Kinect based facial emotion recognition systems use both RGB and depth data for extracting different feature points. In 2013 Seddik et al. [44] recognized facial expressions and mapped them to a 3D face virtual model using Kinects depth and RGB data. Breidt et al. [45] released a specialized 3D morphable model for facial expression analysis and synthesis using noisy RGB-D data from Kinect. Their results showed the potential of using Kinect sensor in facial expression analysis. In 2015, Mao et al. [46] proposed a real-time EFRE method, in which both 2D and 3D features extracted with Kinect are used as features. The emotion classification has been done using support vector machine (SVM) classifiers and the recognition results of 30 consecutive frames are fused by the fusion algorithm based on improved emotional profiles (IEPs). Youssef et al.[47] created a home-made dataset containing 3D data for 14 different persons performing the 6 basic facial expressions. To classify emotions, SVM and k-NN classifiers are utilized. They have achieved 38.8% (SVM) and 34.0% (k-NN) classification accuracy for individuals who did not participate in training of the classifiers and observed 78.6% (SVM) and 81.8% (k-NN) accuracy levels for the cases where they have tested their approach with volunteers who did participate in training. Zhang et al. [48] trained decision tree classifiers and used 3D facial points recorded by Kinect as inputs. The best accuracy reached was 80% for three emotions in only female data with decision tree classification. Recently, in 2017, Tarnowski et al [49] constructed the dataset with six men performing 7 emotional states, hence a total of 256 facial expressions. Then, they performed facial emotion classification by using 6 action units tracked by Kinect v1 sensor as inputs to an artificial neural network.

Deep-learning based approaches are also gaining much popularity due to their computational advantages such as enabling end-to-end learning, without a hand-crafted feature extraction process. From scene understanding to facial expression recognition, CNN has achieved state-of-the-art results. An example of a CNN based FER system is illustrated in Figure 2.7 below.

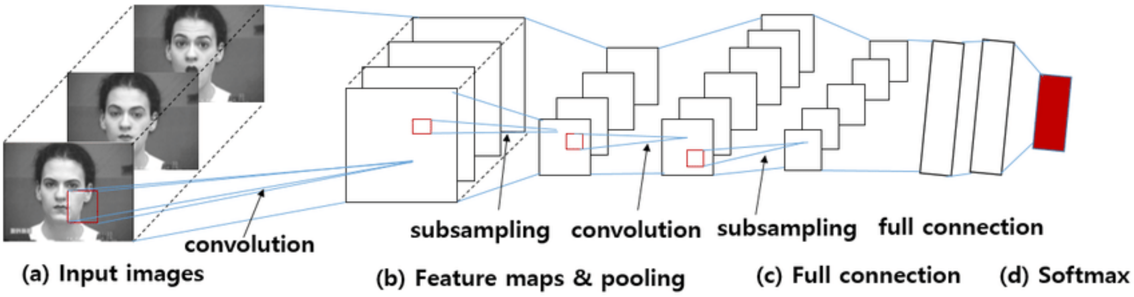


Figure 2.7: Block diagram for a CNN based FER approach (face images are taken from CK+ dataset [4])

The requirement of large data for training is generally one of the barriers to use deep learning methods. As there are many parameters in the model to learn, to avoid the overfitting, the amount of data also has to be very large.

Even though considerable success has been achieved with both deep learning based and conventional methods, there are still a great number of issues remained which deserve further investigation. Some of these problems are listed below:

- A wide range of datasets and superior computing/processing power are demanded.
- A great number of manually collected and labeled datasets are required.
- Large memory is needed.
- Both training and testing processes are time consuming.
- Expertise is required to select appropriate parameters including learning rate, kernel sizes filters, number of neurons and number of layers.
- Even though CNNs work well for various applications, there are several criticism toward CNNs regarding the lack of theory as it needs to rely on trials-and-errors.

TABLE 2.1: Recognition performance of certain implementations with MMI dataset, adapted from [20].

Type	Brief Description of Main Algorithms	Input	Accuracy(%)
Conventional (handcrafted-feature) FER approaches	Sparse representation classifier with LBP features [50]	Still frame	59.18
	Sparse representation classifier with local phase quantization features [51]	Still frame	62.72
	SVM with Gabor wavelet features [52]	Still frame	61.89
	Sparse representation classifier with LBP from three orthagonal planes[53]	Sequence	61.19
	Sparse representation classifier with local phase quantization feature from three orthagonal planes[54]	Sequence	64.11
	Collaborative expression representation CER [55]	Still frame	70.12
	Average		63.20
Deep-learning-based FER approaches	Deep learning of deformable facial action parts [56]	Sequence	63.40
	Joint fine-tunning in deep neural networks [57]	Sequence	70.24
	AU-aware deep networks [58]	Still frame	69.88
	AU-inspired deep networks [59]	Still frame	75.85
	Deeper CNN [60]	Still frame	77.90
	CNN+LSTM with spatio temporal feature representation[61]	Sequence	78.61
	Average		72.65

Chapter 3

Depth Sensor and Dataset Generation

In this chapter the sensor utilized in the experimental part of the thesis and the dataset generation will be detailed.

3.1 Depth Sensor

Depth sensors have gained popularity with the rise in their application fields. AR/VR, gesture and face recognition, mapping, navigation and automation are only some of the application areas in robotics, security, automotive, aviation, entertainment industries.

3D depth sensing techniques have evolved dramatically in the last two decades. There are diverse range of 3D shape acquisition methods. Depth information to the machine can be derived from mainly contact, in which machine and object are physically in contact, and non-contact methods. In various industries including healthcare, aviation, mining etc., non-contact methods have long been used to acquire depth information. To name a few, tomography and sonar machines are the most popular among others. Tomography machine is based on transmissive techniques and uses ionizing radiation to gather the data, whereas sonar relies on reflection of sound waves.

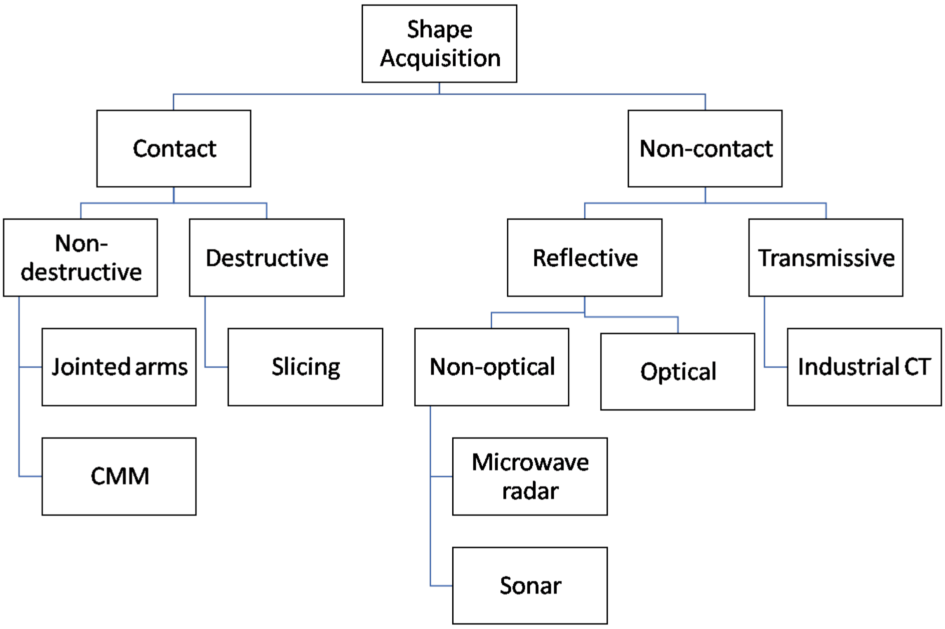


Figure 3.1: Classification of 3D data acquisition techniques [6]

Optical shape acquisition/ depth sensing methods are fairly new compared to other techniques. In depth sensors, stereo, a well-known passive method, structured light and time of flight (ToF), as active methods, are the most widely used principles. As an example, Kinect v1 is based on structured light principle and Kinect v2, which was used in our experiments, is based on time of flight principle.

3.1.1 Stereo

Stereo method is based on observing an object from two different points of view. The distance in pixels between two corresponding points in a pair of stereo images is called disparity. Using disparity maps, depth information can be computed with

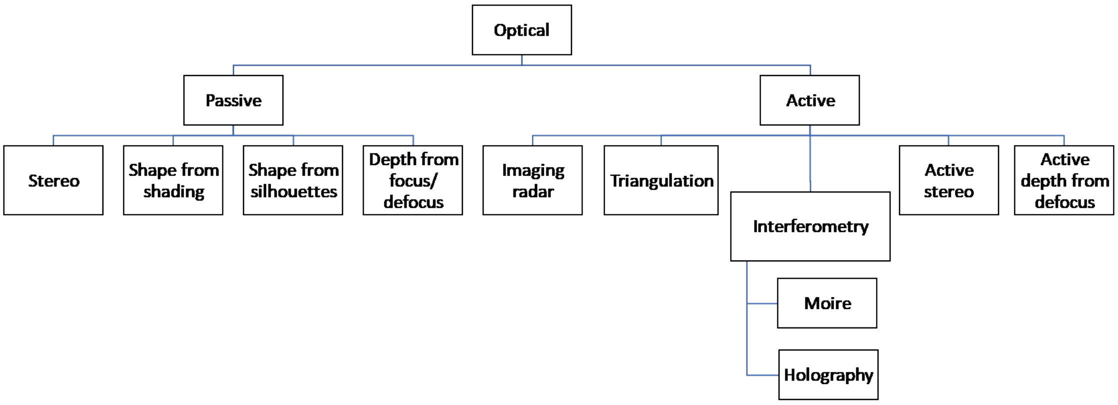


FIGURE 3.2: Classification of optical 3D data acquisition methods [7]

a method called triangulation. Below is the demonstration of geometry of stereo method:

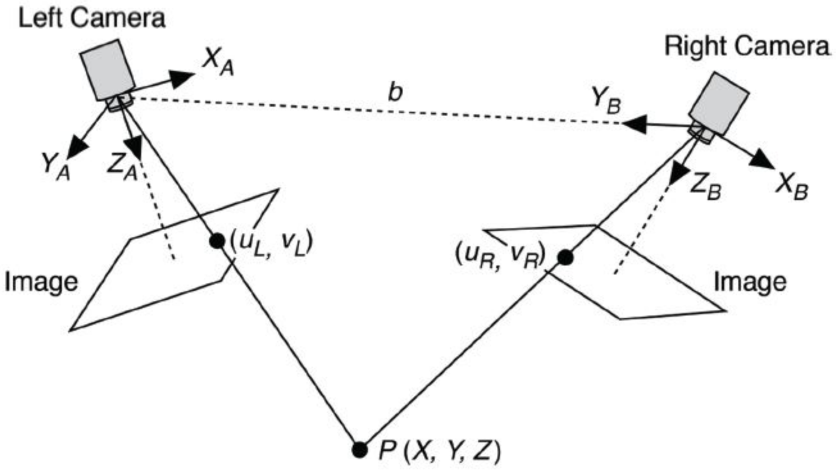


FIGURE 3.3: Stereo vision model [8]

Stereo is one of the oldest techniques. Stereo cameras have been around for more than 100 years. It gained popularity in the last 2 decades with the rise of 3D movie market.

3.1.2 Structured Light

In structured light, an active stereo method, a light pattern is projected on to the object by a projector. The pre-determined pattern is distorted by the object. Another camera in a certain distance from the projector, is used to observe the deformations in the pattern in order to acquire and calculate the depth data. In simpler terms, the pattern projection on close objects are deformed more, on far objects the distortion is less intense.

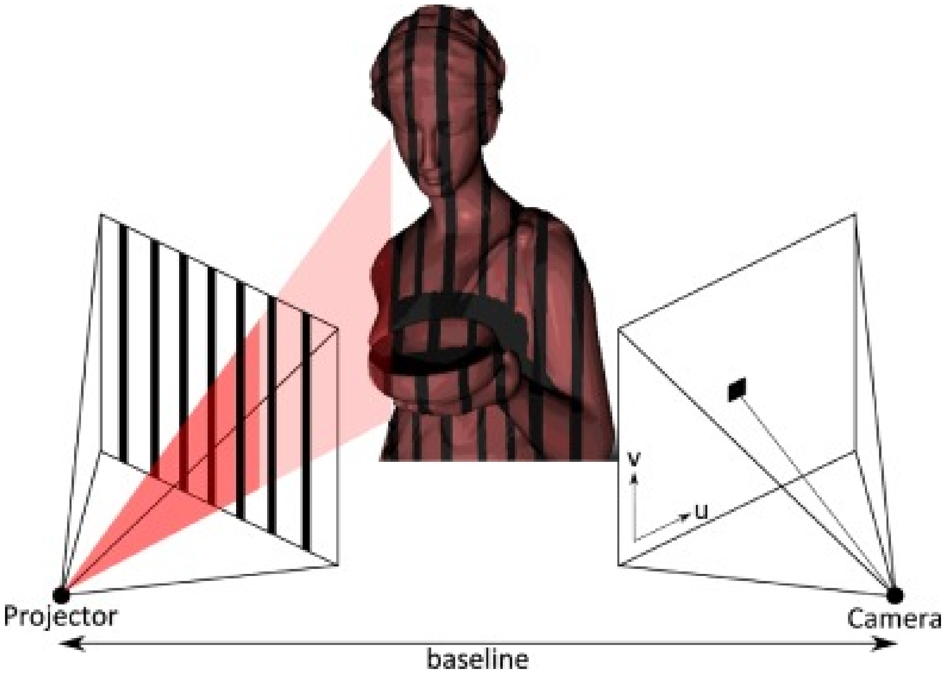


FIGURE 3.4: Structured light approach [9]

Given the intrinsic parameters of the camera, such as the focal length f and the baseline b between the camera and the projector, the depth information of a 2D point (x, y) can be calculated as $d = \frac{bf}{m(x,y)}$ using the disparity value $m(x, y)$. The unit of the disparity $m(x, y)$ is usually given as pixel, thus, a unit changing operation for the focal length also takes place to convert it into pixel units, i.e. $f = \frac{f_x^{metric}}{s_{px}}$, where s_{px} denotes the pixel size.

Kinect v1 which was launched in 2010 by Microsoft, works based on the structured light principle.

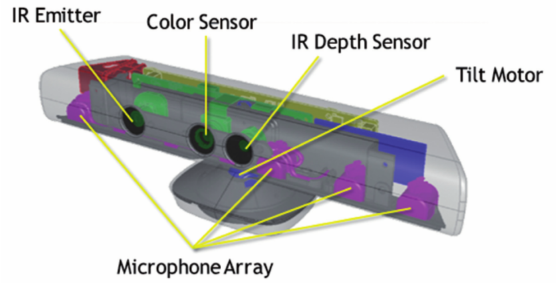


FIGURE 3.5: Construction of structured-light Kinect [10]

The Kinect v1 is made of a color RGB camera, a monochrome NIR camera, and an NIR projector with a laser diode at 850nm wavelength. Based on the disparity between the initial image and the image recorded by IR camera, Kinect uses triangulation method to calculate the distance of the objects. There are variety of light patterns of projections. Striped pattern is the simplest and common case, whereas Kinect v1 makes use of a structured dot pattern in infa-red. Structured light approach is also utilized by Apple in its iPhone X model (see Figure 3.6).

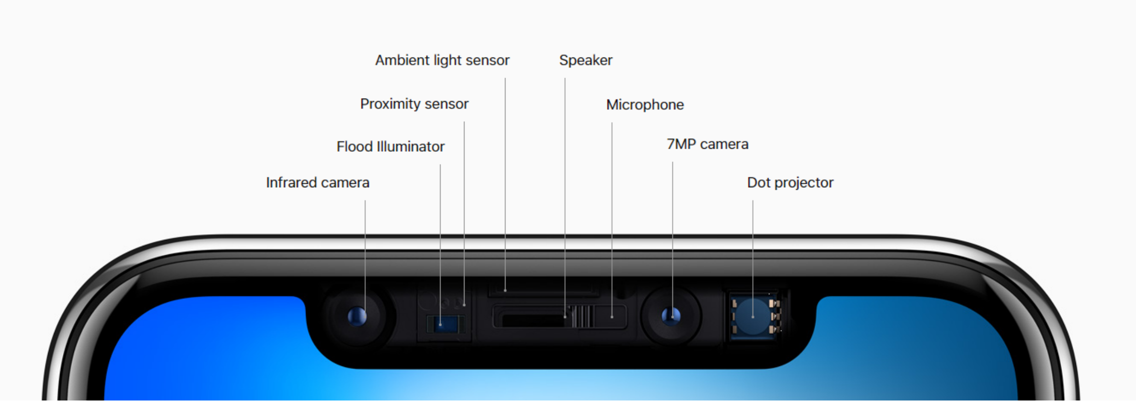


FIGURE 3.6: The components of Apple's TrueDepth camera [11]

3.1.3 Time of Flight

The time of flight (ToF) technology is based on measuring the time difference between emitted light and its return to the sensor after reflection from the object. The time of flight technology is the basis of several range sensing devices and ToF cameras, to name the most popular, Kinect v2. As in many ToF cameras, Kinect v2 also uses Continuous Wave (CW) Intensity Modulation approach. The method is based on continuously projecting intensity modulated periodic light on to the object. The distance between camera and the object causes a delay $\phi[s]$ in the signal which deviates the phase in the periodic light (see Figure 3.7). The deviation of time is observed for every pixel. Given that, speed of light c, the object distance for each pixel can be calculated by the formula $d = \frac{c\phi}{f_m 4\pi}$ where f_m is called modulation frequency.

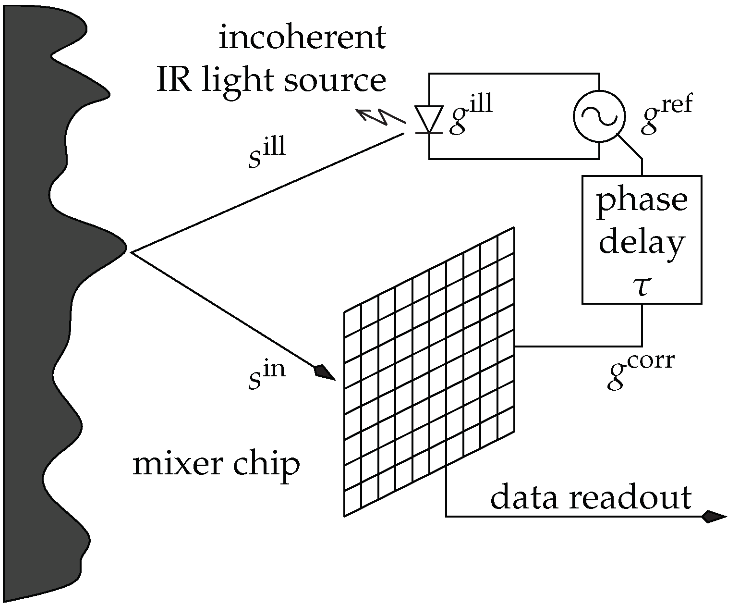


FIGURE 3.7: The time of flight (ToF) phase-measurement principle [9]

3.1.4 Comparison of three techniques

Each technique presented and described above, has various advantages and setbacks. For instance, accuracy of the structured light method is higher than the time of flight, whereas real-time capability and XY resolution is much higher in time-of-flight techniques. Since the sunlight clears out the infra-red light, structured light has very low performance in outdoors. Based on different needs in various applications, a suitable method can be chosen. Below is a brief comparison of different techniques.

TABLE 3.1: The performance comparison of three 3D measuring techniques

3D Measuring Technique	**Stereo**	**Structured Light**	**Time of Flight**
XY Resolution	Scene Dependent	Medium	High
Accuracy	Low	High	Medium
Software Complexity	High	Medium	Low
Real-time Capability	Low	Medium	High
Material Costs	Low	High	Medium
Low-light Performance	Weak	Good	Good
Outdoor Performance	Good	Weak	Medium

In the experiments, Kinect v2, launched by Microsoft in 2014, has been used. In line with our aims, Kinect v2 outweighed other 3D depth sensor cameras in our pre-analysis. The sensor and its components can be seen in Figure 3.8 and Figure 3.9 respectively.

FIGURE 3.8: Kinect v2 sensor

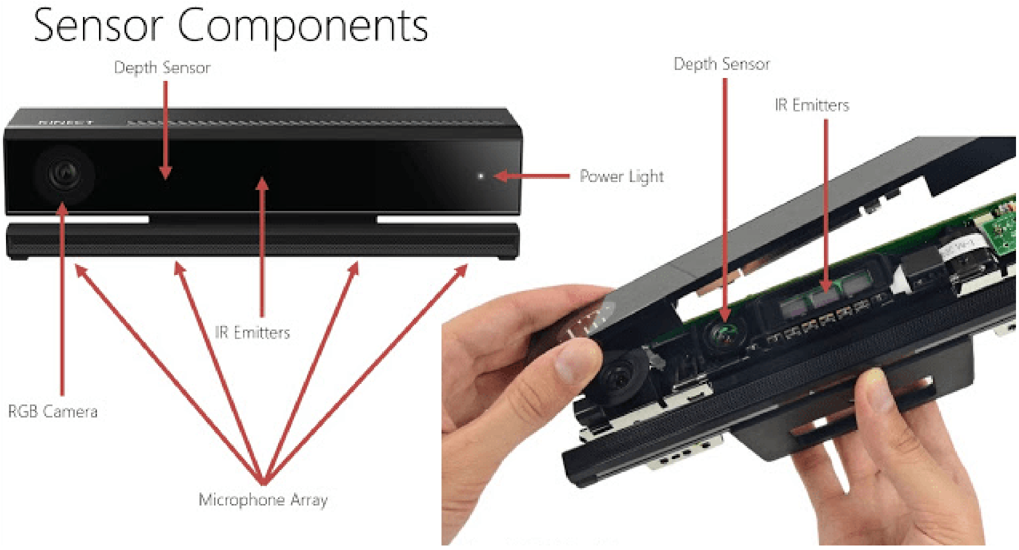


FIGURE 3.9: Sensor components of Kinect v2

Technical features of Kinect v2 sensor is listed below.

TABLE 3.2: Technical features of Kinect v2 sensor

Infrared (IR) Camera Resolution	512 X 424 pixels
RGB Camera Resolution	1920 x 1080 pixels
Field of View	70 x 60 degrees
Frame Rate	30 frames per second from 0.5 to 4.5 m
Operative Measuring Range	between 1.4 mm (@ 0.5 m range)
Object Pixel Size (GSD)	and 12 mm (@ 4.5 m range)

3.2 Action Units from Kinect v1 and Kinect v2

High XY resolution enables an efficient use of high definition face tracking API. In comparison with Kinect v1, Kinect v2 sensor can detect 17 action units (AUs) whereas Kinect v1 can detect only 6. Increase in the number of detected action units, boosts the accuracy of the emotion recognition. The diffence in terms of extracted action units is illustrated in the Figure 3.10 below.

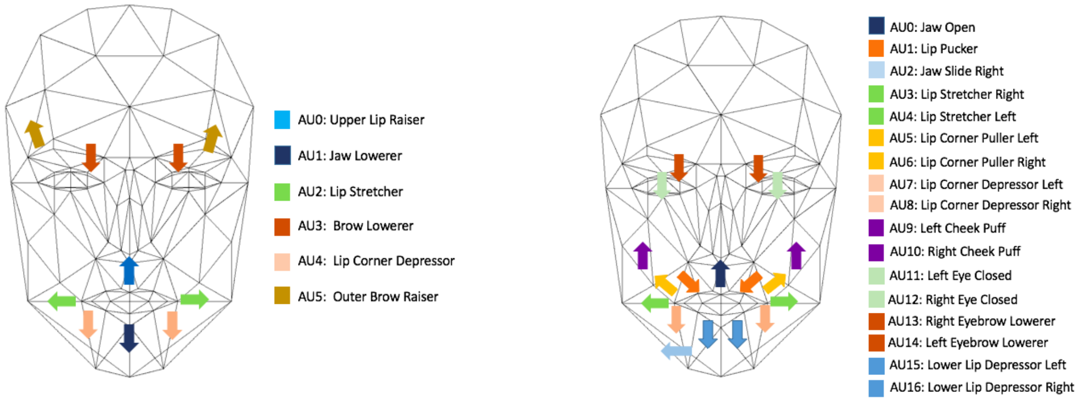


FIGURE 3.10: Visualization of action units (AUs) extracted from Kinect v1 and v2, respectively. Color-coded labels of AUs indicate the position of that specific muscle. The arrow signs are used to illustrate the muscle movement

As it can be seen from the figure 3.10 above, 11 extra action units are mostly situated in the mouth area and therefore they are better at representing emotions that involve movement of muscles around lower face. When it comes to the classification of complex emotions such as disgust and anger, the new set of features plays a crucial role.

3.3 Dataset Generation

The dataset we use in our experiments contains 910 images of both male and female facial expressions. Thirteen volunteers (eight males and five females) who are all graduate students were asked to pose several different facial expressions. In order to obtain more realistic and natural emotional responses instead of just mimicking, an emotion priming experiment was performed by showing volunteers different emotional videos. Each subject was seated at a distance of two meters away from the Kinect sensor and the videos were played with a computer. While they were watching these videos, Kinect v2 was used to record the facial data of subjects.

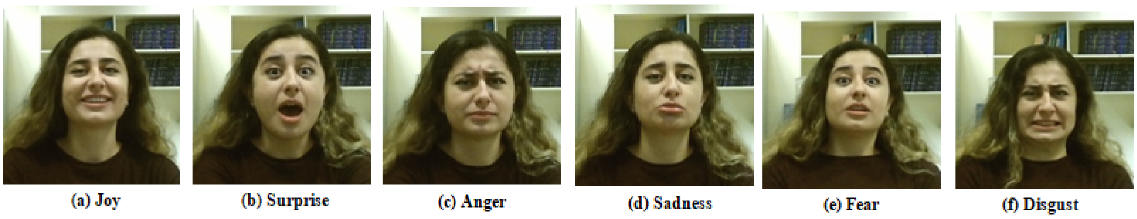


Figure 3.11: Sample facial expression images taken from the dataset

The experiments were conducted in a laboratory at Sabanci University. Each participant took 10 second breaks between emotional states and performed each emotion for a minute. For each emotion, 10 peak frames have been chosen per participant. As a result, 70 frames (10 peak frames × 7 emotions) were collected for each subject. Overall dataset consisted of 910 images (70 frames × 13 participants) facial expressions. Sample images from our dataset are shown in Figure 3.12 below.

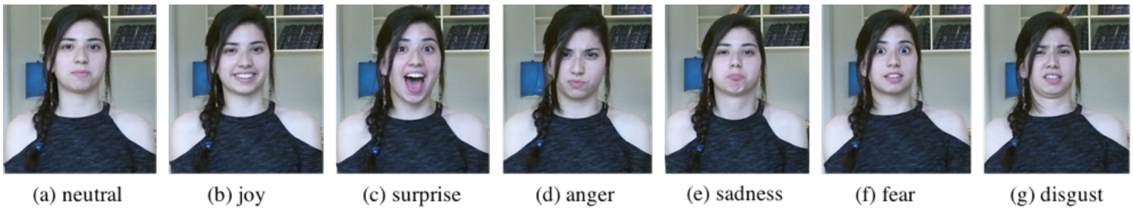


Figure 3.12: Sample facial expression images taken from the dataset

Chapter 4

Feature Extraction and Classification

4.1 Feature Extraction

The feature extraction process is the stage in which the pixel representation of the image is converted into a higher-level representation of shape, motion, color, texture and spatial configuration. The dimensionality of the input space for classification generally decreases after feature extraction.

In facial emotion recognition systems, once the face is detected, the next step is feature extraction where the most appropriate representation of the face for facial emotion recognition is achieved.

The common facial features extraction methods for emotion recognition can be divided into 2 groups: geometric based and appearance based methods (see Figure 4.1).

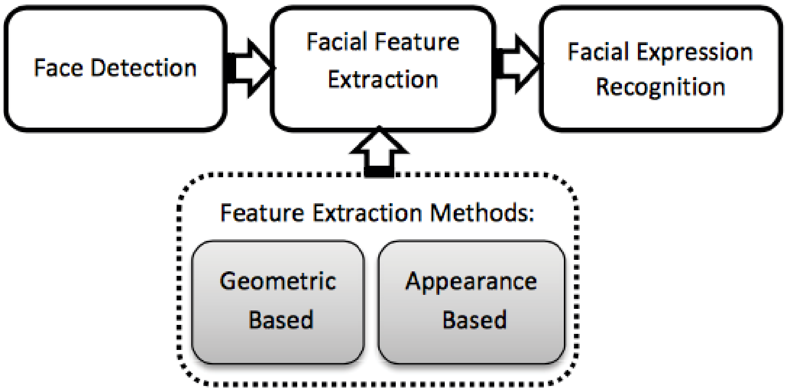

Face Detection
Facial Feature Extraction
Facial Expression Recognition
Feature Extraction Methods:
Geometric Based
Appearance Based

FIGURE 4.1: Facial expressions analysis [12]

4.1.1 Geometric Based Methods

The geometric facial features represent the information regarding the shape and locations of key facial parts such as mouth, eyes, brows and nose. Localization and tracking a dense set of feature points are key steps in the geometric based methods.

Active Appearance Models (AAM) and its variations are one of the most famous and used geometric feature extraction methods [62]. Still, to utilize the geometric feature-based methods, it is needed to achieve accurate and reliable facial feature detection and tracking, which is difficult to achieve in many occasions [63].

4.1.2 Appearance Based Methods

Methods based on appearance features (appearance characteristics) focuses on structure of distinct parts of the face such as muscle shifts due to different emotions including crinkles, contractions, furrows and lumps. A facial expression causes shifts in the position of associated muscle groups which induce wrinkles, distention and so on, resulting with the change in the appearance of local area of the face. Several prominent appearance features used for image classification are Gabor Descriptor

[64] and Local Binary Patterns (LBP) (see Figure 4.2) [65] and Histograms of Oriented Gradient (HOG) Descriptors (see Figure 4.3) [66].

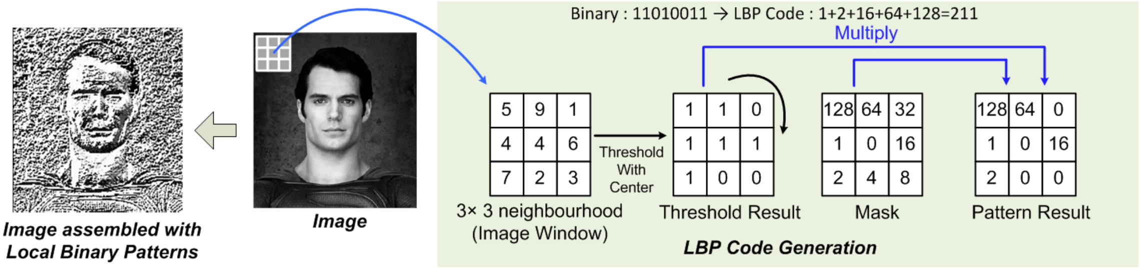


FIGURE 4.2: Local Binary Patterns (LBP) feature extraction process [13]

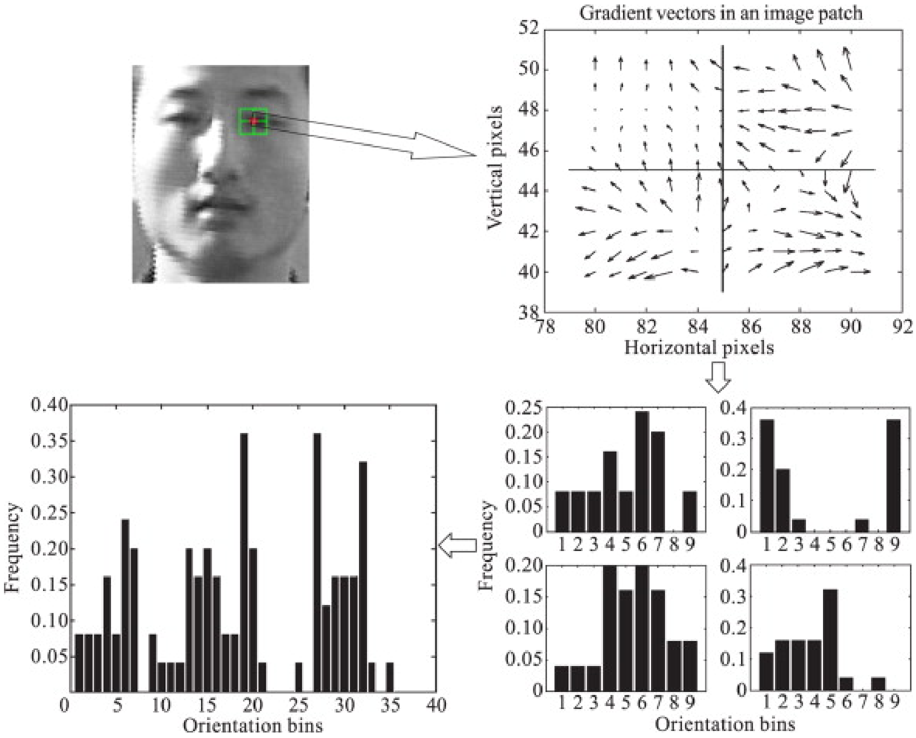


FIGURE 4.3: Histogram Oriented Gradient feature extraction stages[14]

In the approach proposed in this thesis, both geometric and appearance-based features are utilized to develop a system which is more robust to variations in head orientation and light conditions.

Our homemade dataset contains 910 images (70 frames × 13 participants) of facial expressions. Each image is represented with two different feature vectors. In the first representation, seventeen action units (AUs) are extracted with the high definition face tracking API developed by Microsoft. The action unit features coming from each frame can be written in the vector form:

$$a = (AU_0, AU_1, AU_2 ... AU_{16}) \tag{4.1}$$

Thus, each image is represented by a vector of 17 elements. In the second representation, key facial point positions are used as features. Spatial coordinates of these key feature points are written in vector form. Coordinate system attached to the Kinect device is shown in Figure 4.4. The units of measurements are meter and degree for translation and rotation respectively.

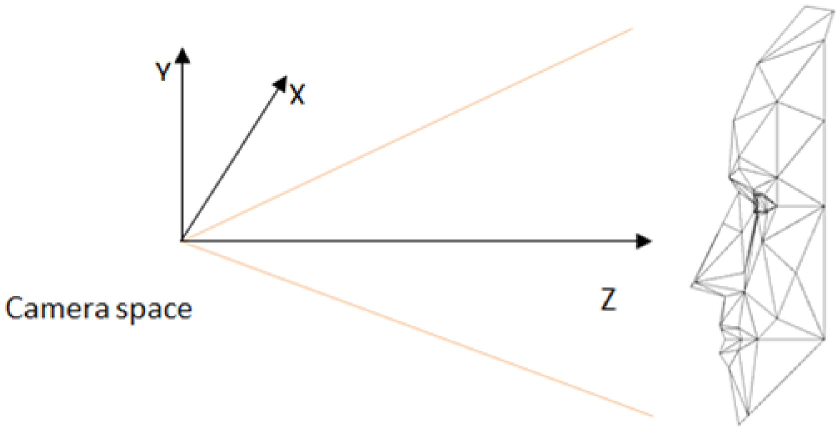


Figure 4.4: Kinect coordinate system

For each frame, 1347 3D facial points are acquired using Face Tracking SDK. However, as it can be seen from Figure 4.5, that not all of these points are strongly related to facial expressions.

In order not to increase the complexity of the training, a dimensionality reduction phase is employed. From these 1347 points, 36 3D points located on eyebrows, chin, mouth, eyes, and some other key positions were selected manually. These 36 key points are also defined with descriptive names on Microsoft website.

For each point, there is 3D information (X, Y, Z). Therefore, the feature point positions (FPP) of each frame can be written in the form of a 108-element vector:

$$b = (X_0, Y_0, Z_0, X_1, Y_1, Z_1, X_{35}, Y_{35}, Z_{35}) \tag{4.2}$$

where $(X_i, Y_i, Z_i), (i = 0, 1, ..., 35)$ are the 3D position coordinates of each 3D key facial point. These 36 key 3D facial points positions are shown in Figure 4.5 whereas descriptions of them are given in Table 4.1.

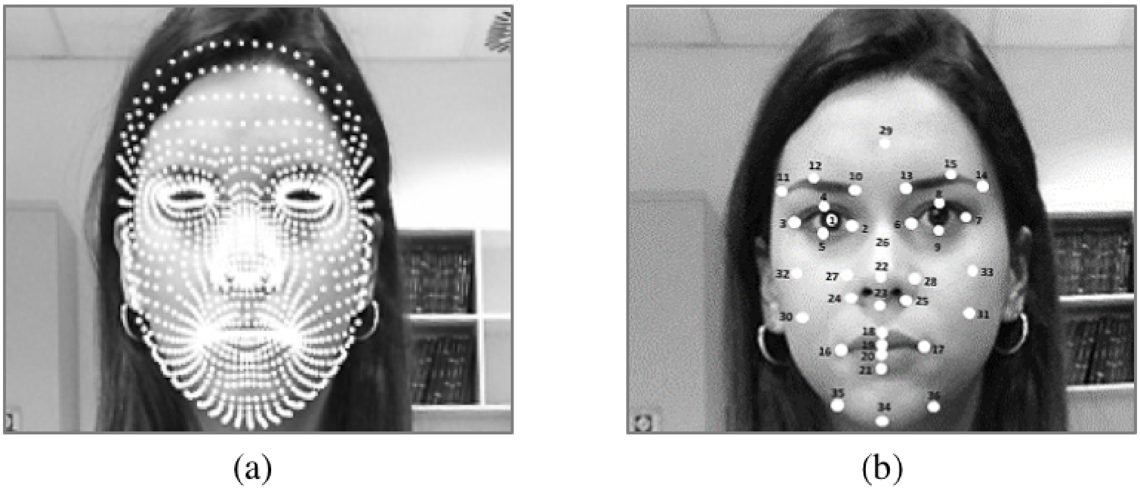


FIGURE 4.5: Facial feature points: (a) The initial 1347 feature points extracted using Kinect face tracking SDK; and (b) the 36 feature points selected as key facial expression features and their enumeration

Face Tracking SDK captures 3D facial points per frame. The sampling frequency of Kinect is 30Hz, thus in one minute 1800 frames can be obtained from one emotion per person.

Table 4.1: List of 3D points and their descriptions

Point no.	Point Description
1	Left eye
2	Inner corner of the left eye
3	Outer corner of the left eye
4	Middle of the top of the left eye
5	Middle of the bottom of the left eye
6	Inner corner of the right eye
7	Outer corner of the right eye
8	Middle of the top of the right eye
9	Middle of the bottom of the right eye
10	Inner left eyebrow
11	Outer left eyebrow
12	Center of the left eyebrow
13	Inner right eyebrow
14	Outer left eyebrow
15	Center of the left eyebrow
16	Left corner of the mouth
17	Right corner of the mouth
18	Middle of the top of the upper lip
19	Middle of the bottom of the lower lip
20	Middle of the top of the lower lip
21	Middle of the bottom of the upper lip
22	Tip of the nose
23	Bottom of the nose
24	Bottom left of the nose
25	Bottom right of the nose
26	Top of the nose
27	Top left of the nose
28	Top right of the nose
29	Center of the forehead
30	Center of the left cheek
31	Center of the right cheek
32	Left cheek bone
33	Right cheek bone
34	Center of the chin
35	Left end of the lower jaw
36	Right end of the lower jaw

4.2 Classification

4.2.1 Artificial Neural Network (ANN)

The human brain is a complex structure which consists of around 100 billion neurons and 1,000 trillion synaptic interconnections. The neurons are responsible for transmitting and processing the information coming from our senses. These electrically excitable cells have three main parts: dendrites, axons, and cell body (soma).

The axon is the long and thin output structure of the neuron. It transfers information to other neuron whereas the dendrites receive the impulse from the synaptic terminals. The general structure of a neuron can be seen in Figure 4.6.

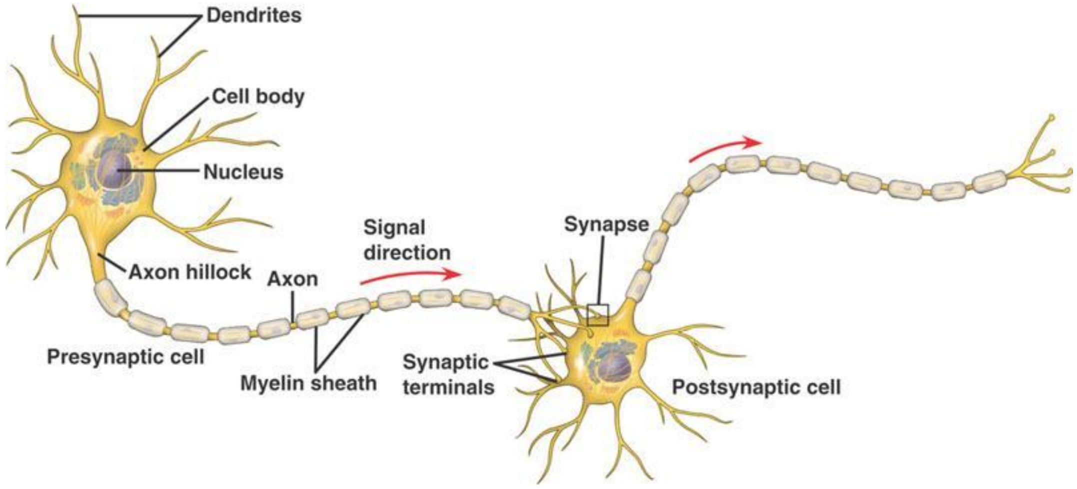


Figure 4.6: Basic neuron parts: dendrites, the cell body, the axon and finally the synapse [15])

An artificial neural network is a computational model inspired by interconnected model of human visual system. It is made of various processing units called artificial neurons.

In the computational model, the signals coming from the synaptic terminals (e.g. x_1,x_2,..,x_n) communicate multiplicatively (e.g. w_0x_0) with the dendrites of the other neuron according to the synaptic strength at that synapse. That synaptic strenght is expressed as weights in the artifical model (e.g. w_{1j},w_{2j},...,w_{nj}).

In the cell body, all the signals coming through the dendrites are summed up. If the final sum is greater than a certain treshold, then the message can be transmitted along the axon to the other synaptic terminal. This structure is modelled with an activation function f as illustrated in Figure 4.7.

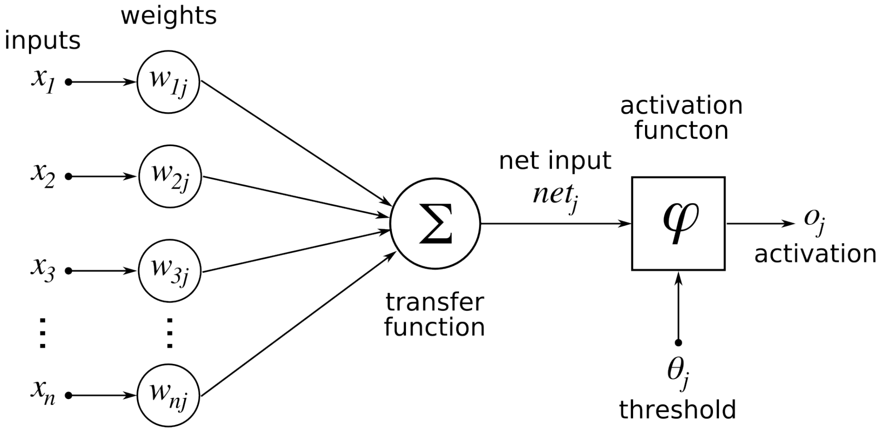


FIGURE 4.7: Simple neural network model [16])

As the activation function is decisive in learning non-linear properties present in the data, the choice of the activation function is pivotal. The choice of the activation function is problem dependent; and there is still a common conception regarding what activation functions work well for prevalent problems. The most common activation functions that can be seen in ANNs are illustrated in Figure 4.8.

4.2.1.1 Network Training

Network training is basically the problem of determining the parameters or so called weights to model the target function. At first, weights are randomly assigned in the model. Based on the initial values of the weights, outputs are calculated. This process is called as forward pass. Then, the error function is computed based on

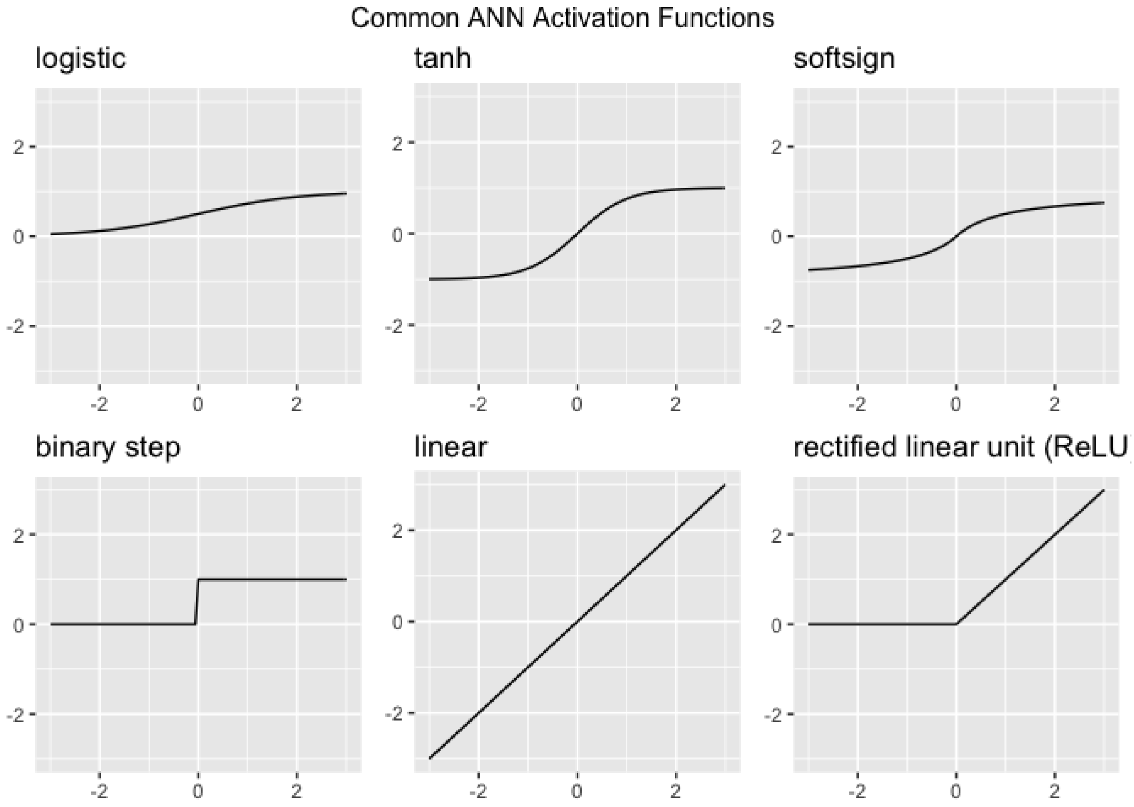


FIGURE 4.8: Common activation functions artificial neural networks

the difference between actual target function and this estimated one. Two common error measures are shown below. Sum-of-squared error function has the form

$$E(w) = \sum_{n=1}^{N} E_n(w) = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{C}(y_k(x_n, w) - t_{nk})^2 \tag{4.3}$$

where N is the number of samples in the dataset and C is the number of classes. The cross-entropy error function is defined as

$$E(w) = \sum_{n=1}^{N} E_n(w) = -\sum_{n=1}^{N}\sum_{k=1}^{C} t_{nk}log(y_k(x_n, w)) \tag{4.4}$$

where t_{nk} denotes the k^{th} entry of the n^{th} target value.

Learning the weights to describe the model requires updating the weights accordingly. Weight update is determined by a parameter optimization algorithm whose aim is to minimize the error function. The error is minimized by differentiating the performance function with respect to the weights. In this process the use of partial derivative is required since each weight is updated individually. Moreover, another scalar parameter called ‘learning rate’ is added to control the step size for weight changes. The weight updates are calculated as follows:

$$\Delta\vec{w} = r * (\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, ..., \frac{\partial E}{\partial w_q}) \tag{4.5}$$

Then, the next step is to update weights according to the selected optimization method. Once the weights are updated, another forward pass takes place. The learning stops when either a pre-defined number of iterations is reached or minimum error rate is achieved.

Some of the optimization techniques that have been used so far are Gradient Descent, Gradient Descent with Momentum, Scaled Conjugate Gradient and BFGS Quasi-Newton. These are the algorithms that are commonly used in neural network training.

Gradient Descent Backpropagation algorithm adjusts the weights in the negative of the gradient (see Figure 4.9), the direction in which the performance function is decreasing most rapidly and calculates the error gradient.

$$w \leftarrow w - \eta\frac{\partial E}{\partial w} \tag{4.6}$$

Gradient Descent with Momentum assists the acceleration of gradients vectors by ignoring the little features in the surface. With the help of momentum, a network can slide through a shallow local minimum [67].

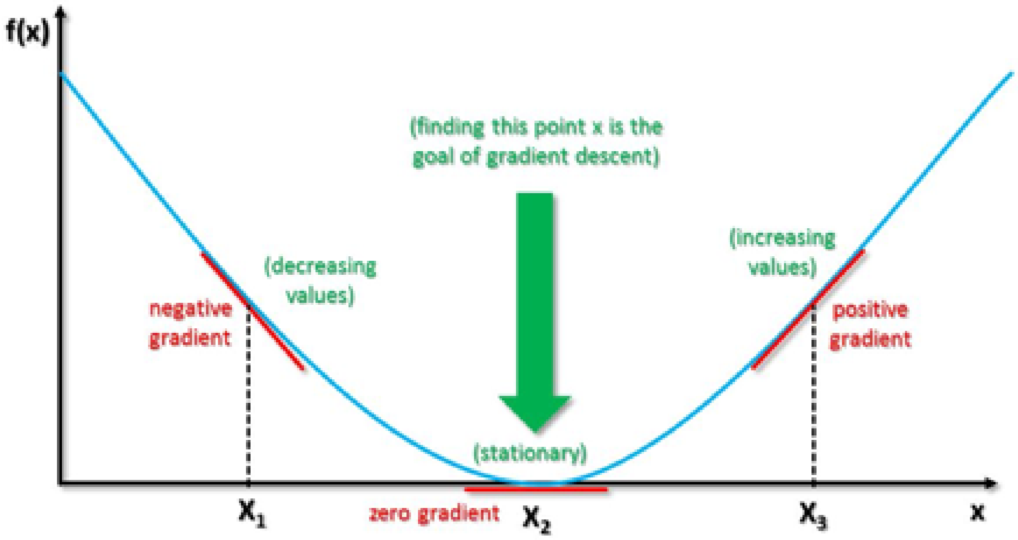


FIGURE 4.9: Illustration of gradient descent algorithm [17]

The conjugate gradient based methods illustrate linear convergence on various problems. Compared to other second order algorithms, it is faster with its step size scaling mechanism. The search paths of the steepest descent and the conjugate gradient methods are illustrated in Figure 4.10.

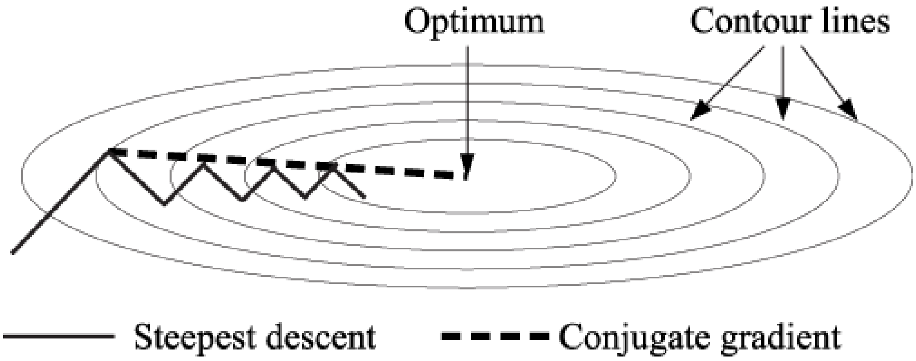


FIGURE 4.10: Search paths of the steepest descent and the conjugate gradient methods on a 2D plane [18]

Another second order algorithm which is commonly used in network training is BFGS quasi-Newton. It is an approximation of Newton's method with class of hill-climbing techniques in search of function's stationary point. It has long been known that this technique works well even for shallow networks.

The Levenberg-Marquardt algorithm works specifically on sum of squared error type of loss functions. Instead of computing the exact Hessian matrix, it calculates the gradient vector and the Jacobian matrix. Due to its dependence on the Jacobian calculation, it fails to work well on big data sets and networks since big Jacobian matrix requires a lot of memory.

Performance comparison between algorithms to train neural networks can be seen below.

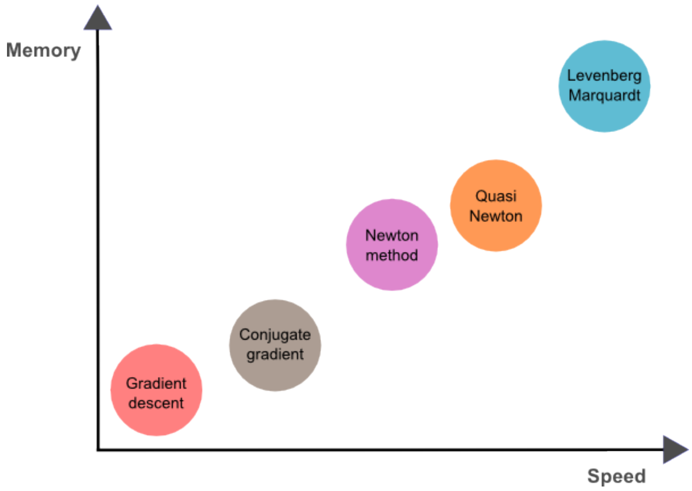


FIGURE 4.11: Comparison of some optimization algorithms in terms of speed and memory [19]

4.2.2 Using ANN to classify emotions

Having extracted two sets of features from each frame, separate neural network classifiers are trained using scaled conjugate gradient backpropagation [68–70]. First classifier has one hidden layer with 10 neurons. In that hidden layer sigmoid action function have been used. Input layer consisted of seventeen action units (AUs) whereas the output was one of the seven emotional states: neutral, joy, surprise, anger, sadness, fear or disgust. The structure of the first neural network can be seen in Figure 4.13.

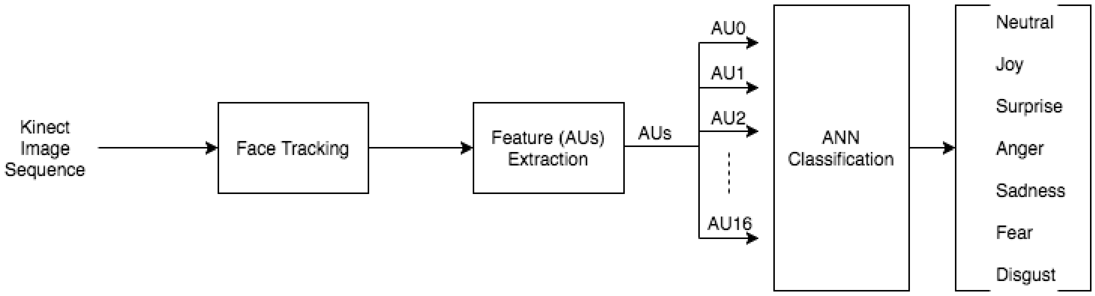


FIGURE 4.12: Schematic representation of the methodology

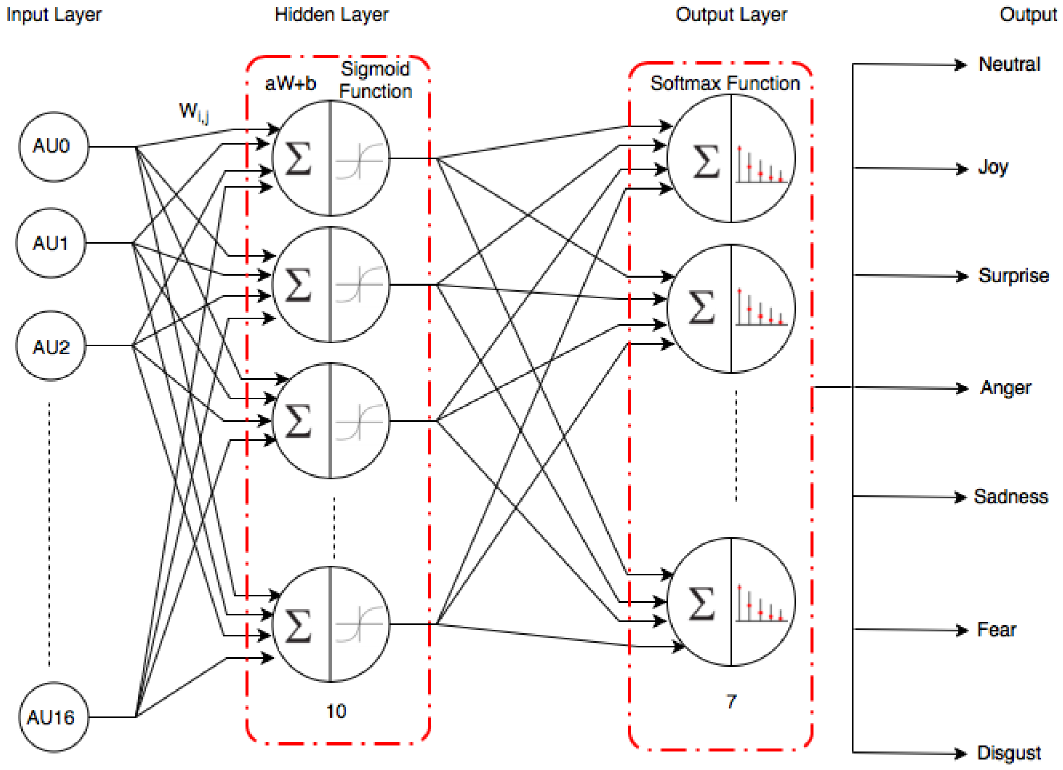


FIGURE 4.13: Architecture of the first neural network classifier

It should be noted that in the output layer of the classifier, a softmax function is utilized which generates class decision probabilities according to the following formula:

$$\sigma(X_i) = \frac{e^{X_i}}{\sum_{j=1}^{C} e^{X_j}} \tag{4.7}$$

where X_i is the input to the softmax function and C is the number of classes.

The second neural network classifier was also trained with scaled conjugate gradient backpropagation. Number of neurons in the hidden layer were increased to 50. The input of the second classifier was 108-element vector consisting of facial feature point positions (FPP) and the output was again one of the seven emotions. To avoid overfitting problem, the overall data was randomly divided into three: training, validation and testing. 70% of data (636 samples) was used for training whereas testing and validation parts were 15% percent each (137 samples). Validation set was used to measure network generalization, and to stop training when generalization stops improving. The structure of the second neural network can be seen in Figure 4.14. .

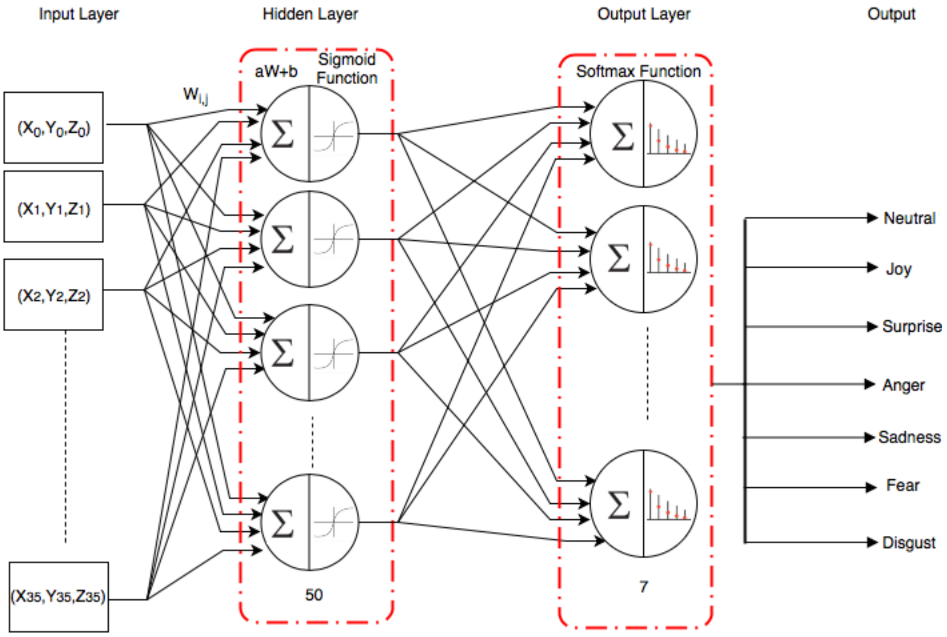


FIGURE 4.14: Architecture of the second neural network classifier

4.3 Ensemble Methods

Recently, ensemble classifiers have gained a lot of research interest as they often induce more accurate and reliable estimates compared to a single model. Variety of papers published by the AI community illustrate that the generalization error is reduced when multiple classifiers are combined [71] [72] [73] [74]. The main reason behind why ensemble methods are very effective is mostly because of a phenomenon called inductive bias [75] [76]. Ensemble methods can reduce the variance error without causing bias error to increase [77] [78]. On several occasions, it is observed that emsembles can reduce bias-error as well [79].

The way of combining individual classifiers can be divided into two categories: simple multiple classifier combinations and meta-combiners. The simple combining methods perform well in the cases where each classifier is reponsible on the same task and has similar success. However, it should be noted that if classifiers are performing unevenly, or the outliers are extreme, then such simple combiners are not suitable. On the other hand, even though, the meta-combiners are theoretically more powerful, they are also vulnerable to overfitting and suffering from long training time. Uniform Voting, Distribution Summation, Bayesian Combination, DempsterShafer, Naive Bayes, Entropy Weighting, Density-based Weighting, DEA Weighting Method and Logarithmic Opinion Pool are the examples of Simple Combining Methods [80].

4.3.1 Using Logarithmic Opinion Pool (LOP)

In the implementation, a softmax activation function [24] has been used at the output layer. The main advantage of using softmax is that it returns the probabilities of each class and the sum of all the probabilities will be equal to one. The output of the network (4.8) is in the form of a $C \times 1$ vector. Each entry of the output vector shows the conditional probability of the label being assigned to the input sample y.

$$O = [p_i(1|y)\ p_i(2|y)...p_i(C|y)]^T \tag{4.8}$$

where C represents the total number of classes, and $i \in \{1,2\}$ represents each classifier.

An ensemble method called Logarithmic Opinion Pool (LOP) [24, 25] is used to combine the decision probabilities coming from each classifier and estimate the final global membership function (4.9)

$$p(c|y) = \frac{1}{Z_{LOP}(y)} \prod_i p_i(c|y)^{w_i} \tag{4.9}$$

where $\sum_i w_i = 1$. Since a uniform distribution is assumed while employing the fusion algorithm, $w_1 = w_2 = 1/2$. $Z(y)$ is a normalization constant which is defined as

$$Z_{LOP}(y) = \sum_c \prod_i p_i(c|y)^{w_i} \tag{4.10}$$

The LOP method treats the outputs of the ensemble members as independent probabilities.The final label of the sample is decided according to (4.11).

$$Label = \operatorname*{argmax}_{c=1..C} p(c|y) \tag{4.11}$$

Chapter 5

Experimental Results

The proposed approach was first tested for subject dependent case in which all the data were randomly divided into training, testing and the validation parts. Training part consisted of 70% of overall data (636 samples) whereas testing and validation parts were 15% percent each (137 samples). Validation set is used for tuning the parameters of the network and minimizing the overfitting by stopping the training when the error on the validation set rises. On the other hand, test set is used for performance evaluation. The results of 5 different training examples are depicted in figures under subject dependent case section.

The performance of the network is further tested using volunteers who were not part of the training data. To measure generalization capabilities of the network, the proposed approach is evaluated through 13 - fold cross validation. The network was trained with samples collected from 12 volunteers and tested with the left out 13th volunteer who was not part of the training data. This procedure is repeated for each subject in turn.

Moreover, in order to analyze the effect of gender in proposed classification approach, an additional test was performed. We divided our dataset into half and used the samples collected from men as training data. We then tested the network with remaining samples coming from our three female volunteers.

5.1 Subject Dependent Case

In subject dependent case, proposed approach is tested for the same volunteers took part in our data generation process. Both networks were trained 50 times using scaled conjugate gradient backpropagation. The average score is calculated by repeating the second experiment 50 times as that network gives the best accuracy without overfitting the data. The average test accuracy of each classifier as well as the fusion accuracy are illustrated in the Table 5.1 below.

Table 5.1: Classification performances

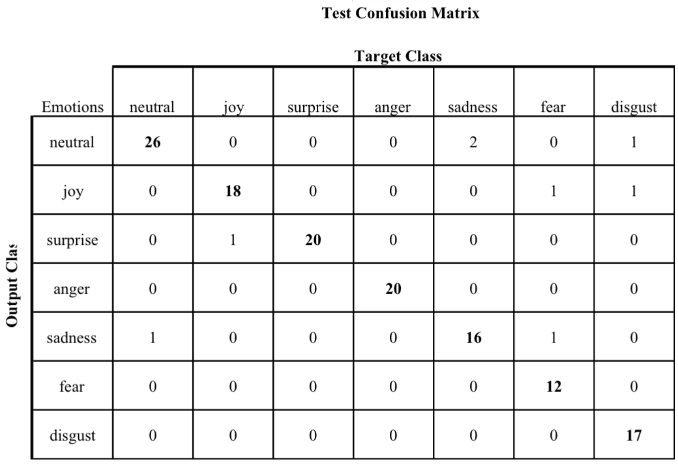
Classifier	**Features**	**Test Accuracy (%)**
NN#1	Action Units(AUs)	92.6
NN#2	3D Feature Points (FPPs)	94.7
Fusion	AU+FPP	97.2

5.1.1 Training Example - 1

- Input of the classifier 1 is a 17x910 matrix, representing 910 samples of 17 elements (action units). It has one hidden layer with 10 neurons and is trained with scaled conjugate gradient backpropagation.

- Input of the classifier 2 is a 108x910 matrix, representing 910 samples of 108 elements (key facial point positions - 36 3D point). It has one hidden layer with 50 neurons and is trained with scaled conjugate gradient backpropagation.

Quantitive results of this example are shown below in Figures 5.1 and 5.2 in the form of a confusion matrix. In the confusion matrix, column index represents the ground truth label whereas the row index represents the predicted label. The numerical values illustrate the number of samples. For instance, in the confusion matrix of the first classifier (see Figure 5.1), row 2 and column 6 has value 1. That means, out of 14 samples labeled as 'fear', the network predicted one sample as 'joy' whereas its actual label was 'fear'.



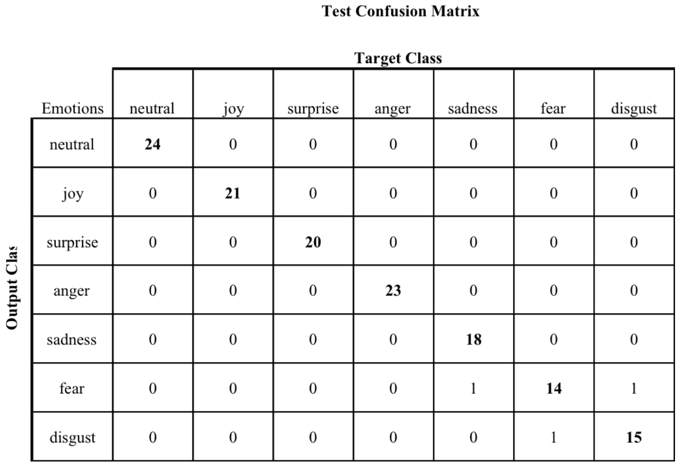
Test Confusion Matrix

Target Class

Output Clas

Emotions	neutral	joy	surprise	anger	sadness	fear	disgust
neutral	**26**	0	0	0	2	0	1
joy	0	**18**	0	0	0	1	1
surprise	0	1	**20**	0	0	0	0
anger	0	0	0	**20**	0	0	0
sadness	1	0	0	0	**16**	1	0
fear	0	0	0	0	0	**12**	0
disgust	0	0	0	0	0	0	**17**

FIGURE 5.1: Test confusion matrix for the first classifier



Test Confusion Matrix

Target Class

Output Clas

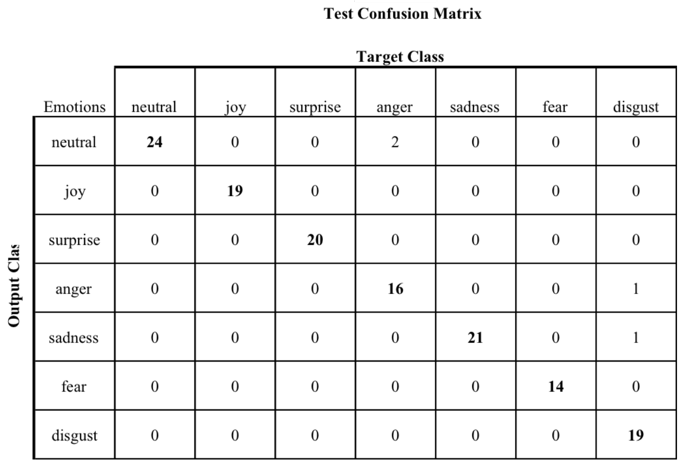
Emotions	neutral	joy	surprise	anger	sadness	fear	disgust
neutral	**24**	0	0	0	0	0	0
joy	0	**21**	0	0	0	0	0
surprise	0	0	**20**	0	0	0	0
anger	0	0	0	**23**	0	0	0
sadness	0	0	0	0	**18**	0	0
fear	0	0	0	0	1	**14**	1
disgust	0	0	0	0	0	1	**15**

FIGURE 5.2: Test confusion matrix for the second classifier

The confusion matrix shows how many predictions was done right within one class and how many of them was done wrong. The diagonal contains the the number of correct classifications whereas the off diagonal includes misclassified samples. Calculation of confusion matrices are important to decide the easiest and the most difficult emotions in term of classification. For the first training example, sadness and fear are most likely to be misclassified.

5.1.2 Training Example - 2

- Classifier 1 has one hidden layer with 20 neurons and is trained with scaled conjugate gradient backpropagation.
- Classifier 2 has one hidden layer with 100 neurons and is trained with scaled conjugate gradient backpropagation.



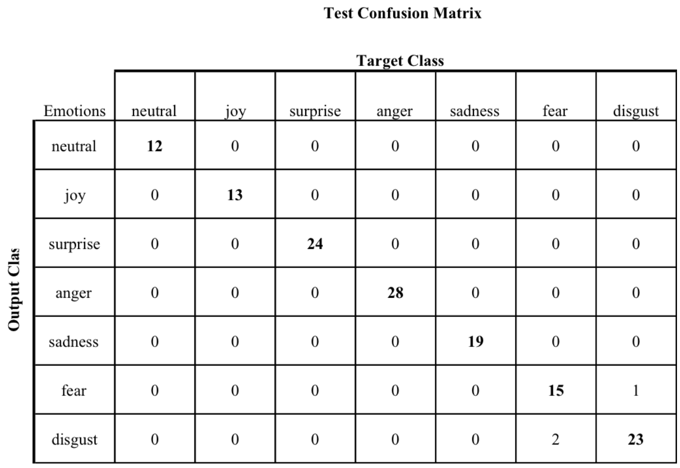
Test Confusion Matrix

Target Class

Output Clas

Emotions	neutral	joy	surprise	anger	sadness	fear	disgust
neutral	**24**	0	0	2	0	0	0
joy	0	**19**	0	0	0	0	0
surprise	0	0	**20**	0	0	0	0
anger	0	0	0	**16**	0	0	1
sadness	0	0	0	0	**21**	0	1
fear	0	0	0	0	0	**14**	0
disgust	0	0	0	0	0	0	**19**

Figure 5.3: Test confusion matrix for the first classifier



Test Confusion Matrix

Target Class

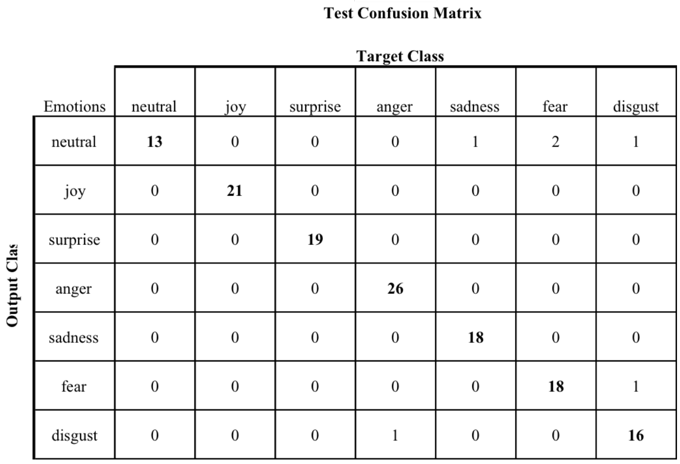
Output Clas

Emotions	neutral	joy	surprise	anger	sadness	fear	disgust
neutral	**12**	0	0	0	0	0	0
joy	0	**13**	0	0	0	0	0
surprise	0	0	**24**	0	0	0	0
anger	0	0	0	**28**	0	0	0
sadness	0	0	0	0	**19**	0	0
fear	0	0	0	0	0	**15**	1
disgust	0	0	0	0	0	2	**23**

FIGURE 5.4: Test confusion matrix for the second classifier

5.1.3 Training Example - 3

- Classifier 1 has one hidden layer with 20 neurons and is trained with scaled conjugate gradient backpropagation.
- Classifier 2 has two hidden layers with 50 neurons each and is trained with scaled conjugate gradient backpropagation.



Test Confusion Matrix

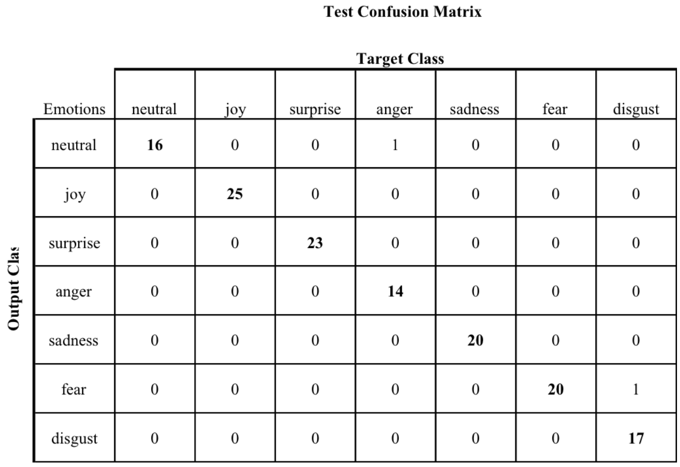
Target Class

Output Clas

Emotions	neutral	joy	surprise	anger	sadness	fear	disgust
neutral	**13**	0	0	0	1	2	1
joy	0	**21**	0	0	0	0	0
surprise	0	0	**19**	0	0	0	0
anger	0	0	0	**26**	0	0	0
sadness	0	0	0	0	**18**	0	0
fear	0	0	0	0	0	**18**	1
disgust	0	0	0	1	0	0	**16**

FIGURE 5.5: Test confusion matrix for the first classifier

Confusion matrix form helps us decide on the emotions with the lowest classification accuracy. According to three subject dependent examples, sadness, fear and



Test Confusion Matrix

Target Class

Output Clas	Emotions	neutral	joy	surprise	anger	sadness	fear	disgust
	neutral	**16**	0	0	1	0	0	0
	joy	0	**25**	0	0	0	0	0
	surprise	0	0	**23**	0	0	0	0
	anger	0	0	0	**14**	0	0	0
	sadness	0	0	0	0	**20**	0	0
	fear	0	0	0	0	0	**20**	1
	disgust	0	0	0	0	0	0	**17**

FIGURE 5.6: Test confusion matrix for the second classifier

disgust are more likely to be misclassified compared to samples labeled as neural, joy, surprise and anger.

5.2 Subject Independent Case

To better evaluate the generalization capabilities of the proposed approach, 13 - fold cross validation is applied through training the network with samples collected from 12 volunteers and testing it with the left out 13th volunteer who was not part of the training data. This procedure is repeated for each subject in turn.

5.2.1 Training Example - 1

- Classifier 1 has one hidden layer with 10 neurons and is trained using conjugate gradient.
- Classifier 2 has one hidden layers with 50 neurons and is trained with scaled conjugate gradient backpropagation.

- The first and the second classifiers' average performances were 63.5% and 53% respectively. When a decision level fusion was applied on both classifiers, the accuracy has increased to 67.5%. The testing results per subject are illustrated in Figure 5.7 below. Acc1 indicates the accuracy of the first classifier whose input is a 17x910 matrix, representing 910 samples of 17 elements (action units) whereas Acc2 represents the second classifier using facial points as features.

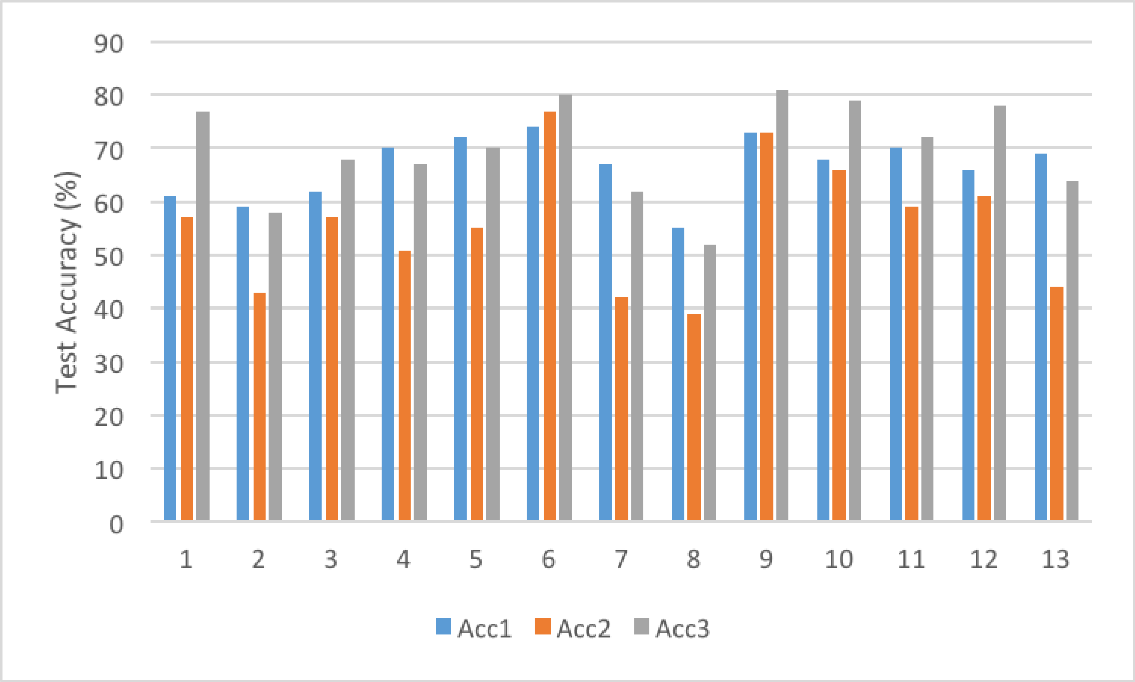


FIGURE 5.7: Accuracy plots for individual classifiers (Acc1 and Acc2) and the combined classifier (Acc3)

Table 5.2: Performances of two classifiers and the combined classifier

	Acc 1(%)	**Acc 2 (%)**	**Acc 3 (%)**
1	55	53	77
2	52	41	50
3	66	53	68
4	67	37	67
5	68	56	70
6	65	70	69
7	60	44	56
8	59	32	52
9	71	72	81
10	66	61	79
11	72	66	72
12	65	59	78
13	60	48	59
Average	63.5	53.2	67.5

Standart deviations for two classifiers and the combined classifier are 5.9, 12.5 and 10.3, respectively.

5.2.2 Training Example - 2

- Classifier 1 has one hidden layer with 20 neurons and is trained using conjugate gradient.
- Classifier 2 has one hidden layers with 100 neurons and is trained with scaled conjugate gradient backpropagation.

- The first and the second classifiers' average performances were 66.6% and 53% respectively for classifying images. When a decision level fusion was applied on both classifiers, the accuracy has increased to 69.8%. The testing results per subject are illustrated in Figure 5.8 below. Acc1 symbolizes the accuracy of the first classifier whose input is a 17x910 matrix, representing 910 samples of 17 elements (action units) whereas Acc2 represents the second classifier using facial points as features.

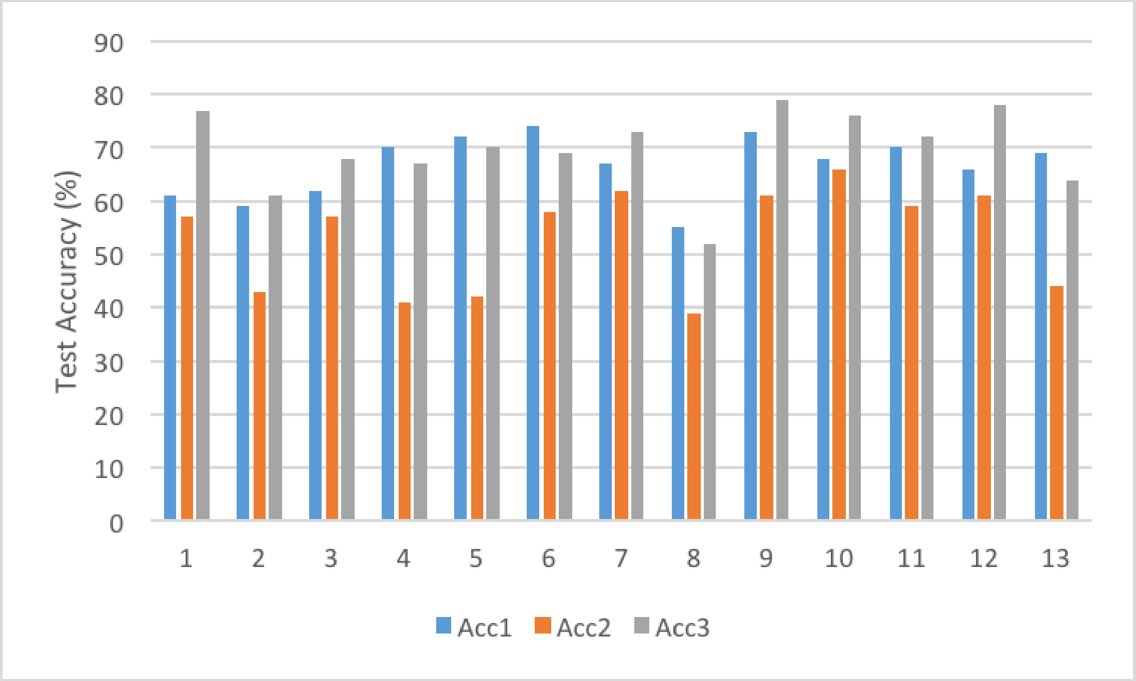


FIGURE 5.8: Accuracy plots for individual classifiers (Acc1 and Acc2) and the combined classifier (Acc3)

Table 5.3: Performances of two classifiers and the combined classifier

	Acc 1(%)	**Acc 2 (%)**	**Acc 3 (%)**
1	61	57	77
2	59	43	58
3	62	57	68
4	70	51	67
5	72	55	70
6	74	77	80
7	67	42	62
8	55	39	52
9	73	73	81
10	68	66	79
11	70	59	72
12	66	61	78
13	69	44	64
Average	66.6	55.6	69.8

Standart deviations for two classifiers and the combined classifier are 5.7, 11.8 and 8.7, respectively.

5.3 Gender-Based Testing

It has been long known that several factors including pose and lighting conditions, gender, age, and facial hair dramatically affect the quality and accuracy of emotion recognition systems. To analyze the effect of gender in proposed classification method, an additional test is applied. First, the samples collected from men were used as training data. Then, network was tested with remaining samples coming from 5 female subjects. For this gender-based test, number of samples and test accuracy is shown in Table 5.4 below.

Table 5.4: Classification accuracy of gender-based test

	Training Data	**Test Data**	**Accuracy (%)**
Data	Male Dataset	Female Dataset	58
Number of Samples	560	350	

Distinct characteristics of female and male face anatomy led to substantial differences between the training and the test data, resulting in a major decrease in the accuracy.

Chapter 6

Conclusion and Future Work

In this thesis, we have presented a facial emotion recognition approach based on the idea of ensemble methods to classify seven different emotional states. Action units and key point feature positions, together with a probabilistic fusion algorithm, enable us to recognize seven basic emotions via facial expressions. We first started by creating our own homemade dataset which consists of 910 samples captured from 13 people. Then, each sample is labeled as neutral, joy, sadness, anger, surprise, fear or disgust. Having extracted two kinds of facial features, action units and feature point positions, separate neural network classifiers are trained with scaled conjugate gradient backpropagation algorithm. To improve the performance of our system decision level fusion is performed. Logarithmic Opinion Pool (LOP) is used as the fusion algorithm. For subject dependent case, the average accuracies using AUs and FPPs were 92.6% and 94.7%, respectively. When fusion algorithm is employed the accuracy has increased to 97.2. It should be noted that even though the accuracy of using FPPs as input is higher than that of using AUs, AUs are only 17- element features, while FPPs are 108-element features.

To further evaluate the performance of the network, the homemade dataset is divided into 13 equal pieces where each 70 samples represents the facial data coming from one subject. Afterward, 13-fold cross validation is performed by training on 12 subjects and testing on the left out 1 subject. This experiment is repeated for each subject in turn. To decide the best network model with less number of parameters and high accuracy along with the shorter execution time, three different experiments are performed. Again, highest accuracy without overfitting is achieved with example-2 where the first and the second networks had one hidden layer with 20 neurons and 100 neurons respectively. While average performances of the first and the second NN classifiers were 66% and 55%, when a decision level fusion was applied the accuracy has increased to 69.8%. However, it should be noted that the fusion algorithm only works well in the cases where the results from both classifiers are close to each other. When one of the classifiers had much less accuracy compared to the other one as occurred in test subjects 2,7 and 8, fusion can give lower classification accuracy.

There might be some limitations derived from the dataset used. It has been known that beside distinctive characteristics of human expressions, lighting of the environment, head orientation of subject and distances from the sensor are highly crucial for Kinect sensor to capture feature points properly. Although, we tried to control head orientation of the volunteers and their distance from the sensor, we did not consider the external disturbances acting on the sensor, such as humidity and temperature. On the other hand, there can be a high intra-subject correlation of FPP feature when the significant decrease on classification accuracy (from subject dependent case to independent case) is considered.

As a future work, we will study CNN based facial emotion recognition to overcome the bottlenecks of hand-crafted features and the limitations of Face Tracking SDK. We will also build a larger facial expression database which contains people with different ages and ethnicity.

Bibliography

[1] M. J. Fehrenbach and S. W. Herring, *Illustrated Anatomy of the Head and Neck-E-Book*. Elsevier Health Sciences, 2015.

[2] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, J. R. Movellan, *et al.*, "Automatic recognition of facial actions in spontaneous expressions.," *Journal of multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[3] M. Rydfalk, "Candide, a parameterized face, report no," *LiTH-ISY-I-866, University of Linkoping, Sweden*, 1987.

[4] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 94–101, IEEE, 2010.

[5] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *sensors*, vol. 18, no. 2, p. 401, 2018.

[6] D. Page, A. Koschan, S. Voisin, N. Ali, and M. Abidi, "3d cad model generation of mechanical parts using coded-pattern projection and laser triangulation systems," *Assembly Automation*, vol. 25, no. 3, pp. 230–238, 2005.

[7] E. Lachat, H. Macher, T. Landes, and P. Grussenmeyer, “Assessment and calibration of a rgb-d camera (kinect v2 sensor) towards a potential use for close-range 3d modeling,” *Remote Sensing*, vol. 7, no. 10, pp. 13070–13097, 2015.

[8] D. Lei, Z. Huang, P. Bai, and F. Zhu, “Experimental research on impact damage of xiaowan arch dam model by digital image correlation,” *Construction and Building Materials*, vol. 147, pp. 168–173, 2017.

[9] H. Sarbolandi, D. Lefloch, and A. Kolb, “Kinect range sensing: Structured-light versus time-of-flight kinect,” *Computer vision and image understanding*, vol. 139, pp. 1–20, 2015.

[10] F. M. CRACOVIENSIA and M. M. DŁUGOSZ, “Structured-light 3d scanner in use to assess the human body posture in physical therapya pilot study,” *Folia medica Cracoviensia*, vol. 54, no. 1, pp. 21–35, 2014.

[11] “iphone x - technical specifications.”

[12] Y.-I. Tian, T. Kanade, and J. F. Cohn, “Recognizing action units for facial expression analysis,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[13] “https://medium.com/@ckyrkou/object-detection-using-local-binary-patterns-50b165658368.”

[14] C. Shu, X. Ding, and C. Fang, “Histogram of the oriented gradient for face recognition,” *Tsinghua Science and Technology*, vol. 16, no. 2, pp. 216–224, 2011.

[15] G. Jordi, A. J. P. RIVERO, and J. G. OJALVO, “Integration of the information in complex neural networks with noise,” 2011.

[16] C. A. de Sousa, “An overview on weight initialization methods for feedforward neural networks,” in *Neural Networks (IJCNN), 2016 International Joint Conference on*, pp. 52–59, IEEE, 2016.

[17] S. Hochreiter, A. S. Younger, and P. R. Conwell, “Learning to learn using gradient descent,” in *International Conference on Artificial Neural Networks*, pp. 87–94, Springer, 2001.

[18] Y.-J. Chan and D. Ewins, “The amplification of vibration response levels of mistuned bladed disks: its consequences and its distribution in specific situations,” *Journal of Engineering for Gas Turbines and Power*, vol. 133, no. 10, p. 102502, 2011.

[19] “https://www.neuraldesigner.com/blog/5-algorithms-to-train-a-neural-network.”

[20] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, “Recurrent neural networks for emotion recognition in video,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 467–474, ACM, 2015.

[21] A. Mehrabian, “Communication without words,” *Communication theory*, pp. 193–200, 2008.

[22] G. S. Shergill, A. Sarrafzadeh, O. Diegel, and A. Shekar, “Computerized sales assistants: The application of computer technology to measure consumer interest-a conceptual framework,” 2008.

[23] “Microsoft sdk for face tracking documentation : http://msdn.microsoft.com/enus/library/jj130970.aspx.”

[24] G. E. Hinton, “Products of experts,” 1999.

[25] A. Smith, T. Cohn, and M. Osborne, "Logarithmic opinion pools for conditional random fields," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 18–25, Association for Computational Linguistics, 2005.

[26] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns for spatial-spectral classification of hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3681–3693, 2015.

[27] C. Darwin and P. Prodger, *The expression of the emotions in man and animals.* Oxford University Press, USA, 1998.

[28] P. Ekman, "Universals and cultural differences in facial expressions of emotion.," in *Nebraska symposium on motivation*, University of Nebraska Press, 1971.

[29] W. V. Friesen, "Cultural differences in facial expressions in a social situation: an experimental test of the concept of display rules.," 1973.

[30] P. Ekman and W. Friesen, "Facial action coding system: A technique for the measurement of facial action," *Manual for the Facial Action Coding System*, 1978.

[31] P. Ekman, "Facial action coding system (facs)," *A human face*, 2002.

[32] C. E. Izard, . , and M. Weiss, *Maximally discriminative facial movement coding system.* University of Delaware, instructional resources Center, 1979.

[33] T. K. Choudhury, *FaceFacts: study of facial features for understanding expression.* PhD thesis, Massachusetts Institute of Technology, 1999.

[34] S. Kawato and J. Ohya, "Real-time detection of nodding and head-shaking by directly detecting and tracking the" between-eyes"," in *Automatic Face and*

Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, pp. 40–45, IEEE, 2000.

[35] J. J.-J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li, "Detection, tracking, and classification of action units in facial expression," *Robotics and Autonomous Systems*, vol. 31, no. 3, pp. 131–146, 2000.

[36] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[37] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pp. 149–149, IEEE, 2006.

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

[39] P. Wang, F. Barrett, E. Martin, M. Milonova, R. E. Gur, R. C. Gur, C. Kohler, and R. Verma, "Automated video-based facial expression analysis of neuropsychiatric disorders," *Journal of neuroscience methods*, vol. 168, no. 1, pp. 224–238, 2008.

[40] S. Happy, A. George, and A. Routray, "A real time facial expression classification system using local binary patterns," in *Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on*, pp. 1–5, IEEE, 2012.

[41] M. H. Siddiqi, R. Ali, A. M. Khan, Y.-T. Park, and S. Lee, "Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields," *IEEE Transactions on Image Processing*, vol. 24, no. 4, pp. 1386–1398, 2015.

[42] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz, "Framework for reliable, real-time facial expression recognition for low resolution images," *Pattern Recognition Letters*, vol. 34, no. 10, pp. 1159–1168, 2013.

[43] D. Ghimire, S. Jeong, J. Lee, and S. H. Park, "Facial expression recognition based on local region specific features and support vector machines," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 7803–7821, 2017.

[44] B. Seddik, H. Maamatou, S. Gazzah, T. Chateau, and N. E. B. Amara, "Unsupervised facial expressions recognition and avatar reconstruction from kinect," in *Systems, Signals & Devices (SSD), 2013 10th International Multi-Conference on*, pp. 1–6, IEEE, 2013.

[45] M. Breidt, H. H. Biilthoff, and C. Curio, "Robust semantic analysis by synthesis of 3d facial motion," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 713–719, IEEE, 2011.

[46] Q.-r. Mao, X.-y. Pan, Y.-z. Zhan, and X.-j. Shen, "Using kinect for real-time emotion recognition via facial expressions," *Frontiers of Information Technology & Electronic Engineering*, vol. 16, no. 4, pp. 272–282, 2015.

[47] A. E. Youssef, S. F. Aly, A. S. Ibrahim, and A. L. Abbott, "Auto-optimized multimodal expression recognition framework using 3d kinect data for asd therapeutic aid," *International Journal of Modeling and Optimization*, vol. 3, no. 2, p. 112, 2013.

[48] Z. Zhang, L. Cui, X. Liu, and T. Zhu, "Emotion detection using kinect 3d facial points," in *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*, pp. 407–410, IEEE, 2016.

[49] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," *Procedia Computer Science*, vol. 108, pp. 1175–1184, 2017.

[50] M.-W. Huang, Z.-w. Wang, and Z.-L. Ying, “A new method for facial expression recognition based on sparse representation plus lbp,” in *Image and Signal Processing (CISP), 2010 3rd International Congress on*, vol. 4, pp. 1750–1754, IEEE, 2010.

[51] Z. Wang and Z. Ying, “Facial expression recognition based on local phase quantization and sparse representation,” in *Natural Computation (ICNC), 2012 Eighth International Conference on*, pp. 222–225, IEEE, 2012.

[52] S. Zhang, X. Zhao, and B. Lei, “Robust facial expression recognition via compressive sensing,” *Sensors*, vol. 12, no. 3, pp. 3747–3761, 2012.

[53] G. Zhao and M. Pietikainen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[54] B. Jiang, M. F. Valstar, and M. Pantic, “Action unit detection using sparse appearance descriptors in space-time video volumes,” in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 314–321, IEEE, 2011.

[55] S. H. Lee, W. J. Baddar, and Y. M. Ro, “Collaborative expression representation using peak expression and intra class variation face images for practical subject-independent emotion recognition in videos,” *Pattern Recognition*, vol. 54, pp. 52–67, 2016.

[56] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, “Deeply learning deformable facial action parts model for dynamic expression analysis,” in *Asian conference on computer vision*, pp. 143–157, Springer, 2014.

[57] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, “Joint fine-tuning in deep neural networks for facial expression recognition,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2983–2991, 2015.

[58] M. Liu, S. Li, S. Shan, and X. Chen, "Au-aware deep networks for facial expression recognition.," in *FG*, pp. 1–6, 2013.

[59] M. Liu, S. Li, S. Shan, and X. Chen, "Au-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, 2015.

[60] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pp. 1–10, IEEE, 2016.

[61] D. H. Kim, W. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Transactions on Affective Computing*, 2017.

[62] G. J. Edwards, T. F. Cootes, and C. J. Taylor, "Face recognition using active appearance models," in *European conference on computer vision*, pp. 581–595, Springer, 1998.

[63] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[64] C. Liu and H. Wechsler, "Independent component analysis of gabor features for face recognition," *IEEE transactions on Neural Networks*, vol. 14, no. 4, pp. 919–928, 2003.

[65] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[66] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.

[67] B. Sharma and K. Venugopalan, “Comparison of neural network training functions for hematoma classification in brain ct images,” *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 16, no. 1, pp. 31–35, 2014.

[68] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, vol. 49. NBS Washington, DC, 1952.

[69] Y. Notay, “Flexible conjugate gradients,” *SIAM Journal on Scientific Computing*, vol. 22, no. 4, pp. 1444–1460, 2000.

[70] A. V. Knyazev and I. Lashuk, “Steepest descent and conjugate gradient methods with variable preconditioning,” *SIAM Journal on Matrix Analysis and Applications*, vol. 29, no. 4, pp. 1267–1280, 2007.

[71] P. Domingos, “Using partitioning to speed up specific-to-general rule induction,” in *Proceedings of the AAAI-96 Workshop on Integrating Multiple Learned Models*, pp. 29–34, Citeseer, 1996.

[72] J. R. Quinlan *et al.*, “Bagging, boosting, and c4. 5,” in *AAAI/IAAI, Vol. 1*, pp. 725–730, 1996.

[73] E. Bauer and R. Kohavi, “An empirical comparison of voting classification algorithms: Bagging, boosting, and variants,” *Machine learning*, vol. 36, no. 1-2, pp. 105–139, 1999.

[74] D. Opitz and R. Maclin, “Popular ensemble methods: An empirical study,” *Journal of artificial intelligence research*, vol. 11, pp. 169–198, 1999.

[75] S. Geman, E. Bienenstock, and R. Doursat, “Neural networks and the bias/-variance dilemma,” *Neural computation*, vol. 4, no. 1, pp. 1–58, 1992.

[76] T. M. Mitchell *et al.*, “Machine learning. wcb,” 1997.

[77] K. Tumer and J. Ghosh, “Error correlation and error reduction in ensemble classifiers,” *Connection science*, vol. 8, no. 3-4, pp. 385–404, 1996.

[78] K. M. Ali and M. J. Pazzani, “Error reduction through learning multiple descriptions,” *Machine Learning*, vol. 24, no. 3, pp. 173–202, 1996.

[79] P. Bartlett and J. Shawe-Taylor, “Generalization performance of support vector machines and other pattern classifiers,” *Advances in Kernel methodssupport vector learning*, pp. 43–54, 1999.

[80] L. Rokach, “Ensemble methods for classifiers,” in *Data mining and knowledge discovery handbook*, pp. 957–980, Springer, 2005.