IDENTIFICATION OF FACIAL EXPRESSIONS USING ARTIFICIAL NEURAL NETWORKS (ANN)

A MASTER'S THESIS

In

Electrical & Electronics Engineering

Atılım University

by

MİNE ALTINAY GÜNLER

JULY 2012

IDENTIFICATION OF FACIAL EXPRESSIONS USING ARTIFICIAL NEURAL NETWORKS (ANN)

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

OF

ATILIM UNIVERSITY

BY

MİNE ALTINAY GÜNLER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

THE DEPARTMENT OF ELECTRICAL & ELECTRONICS ENGINEERING

JULY 2012

Approval of the Graduate School of Natural and Applied Sciences, Atılım University.

Prof. Dr. Ibrahim Akman

Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Elif Aydın

Head of Department

This is to certify that we have read the thesis “IDENTIFICATION OF FACIAL EXPRESSIONS USING ARTIFICIAL NEURAL NETWORKS (ANN)” submitted by Mine Altınay Günler and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Hakan Tora

Supervisor

Examining Committee Members

Assoc. Prof. Dr. Fuad Aliew
Mechatronics Eng. Dept., Atılım Uni.

Asst. Prof. Dr. A. Buğra Koku
Mechanical Eng. Dept., METU

Asst. Prof. Dr. Hakan Tora
Electrical and Electronics Eng. Dept., Atılım Uni. ____________________

Asst. Prof. Dr. Kasım Öztoprak
Computer Eng. Dept., Karatay Uni. ____________________

Asst. Prof. Dr. M. Efe Özbek
Electrical and Electronics Eng. Dept., Atılım Uni. ____________________

Date: July 16, 2012

I declare and guarantee that all data, knowledge and information in this document has been obtained, processed and presented in accordance with academic rules and ethical conduct. Based on these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Mine Altınay Günler

Signature:

ABSTRACT

IDENTIFICATION OF FACIAL EXPRESSIONS USING ARTIFICIAL NEURAL NETWORKS (ANN)

Günler, Mine Altınay

M.S., Electrical - Electronics Engineering Department

Supervisor: Asst. Prof. Dr. Hakan Tora

July 2012, 57 pages

Facial expressions are a kind of nonverbal communication. They carry the state of emotion of a person. Automatic facial expression identification analysis became popular research area nowadays. It can be used in many areas such that physiology, education, murder squad, analysis of tendency to crime to get a clue about mental signals of a person. This thesis study proposes three different approaches with different methods to identify facial expressions based on artificial neural networks (ANN). Firstly, a tree based neural network structure is proposed. Secondly, hidden layer outputs are used for emotion classification. Finally, a facial features based system is designed. Each proposed methods are implemented by using Matlab and each is capable of identifying smile, angry and scream expressions of a person successfully.

Keywords: Multilayer neural network, facial expressions, principle component analysis, image enhancement, color spaces.

ÖZ

YAPAY SİNİR AĞLARI (YSA) KULLANARAK YÜZ İFADELERİNİ TANIMA

Günler, Mine Altınay

Yüksek Lisans, Elektrik - Elektronik Mühendisliği Bölümü

Tez Yöneticisi: Yrd. Doç. Dr. Hakan Tora

Temmuz 2012, 57 sayfa

Yüz ifadeleri sözsüz iletişimin bir türüdür. Kişinin duygu durumunu barındırırlar. Yüz ifadelerini otomatik olarak analiz etme günümüzde popular bir araştırma alanıdır. Psikoloji, eğitim, cinayet masası, suça eğilim analizi gibi çeşitli alanlarda kişinin zihinsel sinyalleri hakkında ipucu elde etmek için kullanılır. Bu tez çalışması duygu tanıma analizi için yapay sinir ağları (YSA) tabanlı üç değişik yaklaşım önermektedir. İlk olarak, ağaç tabanlı sinir ağları yapısı önerilmiştir. İkinci olarak, duygu sınıflandırılması için gizli katman çıktıları kullanılmıştır. Son olarak, yüz özellikleri tabanlı bir sistem tasarlanmıştır. Önerilen her bir metot Matlab kullanılarak oluşturulmuştur ve her biri gülen, sinirli ve bağıran yüz ifadelerini başarılı bir şekilde tanıyabilmektedir.

Anahtar Kelimeler: Çok katmanlı sinir ağları, yüz ifadeleri, temel bileşen analizi, görüntü iyileştirme, renk uzayları.

To My Parents

To memory of my grandmother, who left us on May 16, and my grandfather

What does the reader do when he wishes to see in what the precise likeness or difference of two objects lies? He transfers his attention as rapidly as possible, backwards and forwards, from one to the other. The rapid alteration of consciousness shakes out, as it were, the points of difference or agreement, which would have slumbered forever unnoticed if the consciousness of the objects compared had occurred at widely distant periods of time. What does the scientific man do when he searches for the reason or law embedded in a phenomenon? He deliberately accumulates all the instances he can find which have any analogy to that phenomenon; and by simultaneously filling his mind with them all, he frequently succeeds in detaching from the collection the peculiarity which he was unable to formulate in the one alone; even though that one had been preceded in his former experience by all those with which he now at once confronts it.

W. James, "Psychology: The Briefer Course ", 1890

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Asst. Prof. Dr. Hakan Tora for his guidance, friendly attitude, encouragement and advice at each step of this study and making me interested in neural network applications. I also would like to thank Asst. Prof. Dr. Kasım Öztoprak for his friendly attitude and encouragement.

I am grateful to have had my fiancé, M.Sc. Computer Engineer, Fikret Pirim by my side. Special thanks to him for his support, patience, sensibility, encourage, motivation and great love.

I owe my deepest gratitude to my parents Asuman Serap and Muzaffer Günler. Beside their endless love, trust and support, I would like to express my special thanks to my mother for bringing me fruits and nuts and my father massaging my shoulders throughout the nights I worked a lot.

I would like to express my sincere thanks to Nilay and Ali Pirim for their faith, encouragement and glory for me.

Despite of the thousands of kilometers thanks Naz Bacgeroğlu who does not grudge her support and motivation from me.

I would like to thank to my friends in my company Vakıflar Bankası EBİS for their friendship, encouragements and smiling faces.

Finally, I would like to express my sincere gratitude to my competent engineer Ali Ümit Zeren who supported me by giving permission to attend master courses despite of the adverse conditions in my company.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2D	-	2 Dimensional
3D	-	3 Dimensional
AR	-	Aleix’ and Robert’s Database
AU	-	Action Units
BP	-	Backpropagation
BWI	-	Black and White Image
C1	-	Neutral
C2	-	Smile
C3	-	Angry
C4	-	Scream
C_b	-	Blue Difference
C_r	-	Red Difference
CMU-PIE	-	Carnegie Mellon University - Pose, Illumination and Expression Database
CRT	-	Cathode Ray Tube
FACS	-	Facial Action Units
GSI	-	Gray Scale Image
HE	-	Histogram Equalization
HEBWI	-	Histogram Equalized Black and White Image
HEGSI	-	Histogram Equalized Gray Scale Image

HMM	-	Hidden Markov Model
HSV	-	Hue, Saturation and Value
JAFFE	-	Japanese Female Facial Expressions Database
LDA	-	Linear Discriminant Analysis
MLP	-	Multilayer Perceptron
ANN	-	Artificial Neural Network
PCA	-	Principle Component Analysis
RBF	-	Radial Basis Function
RF	-	Radio Frequency
RGB	-	Red, Green and Blue
SVM	-	Support Vector Machine
TV	-	Television
YUV	-	Luma and Chrominance

CHAPTER 1

MOTIVATION

William James stated that “the bodily changes follow directly the perception of the exciting fact, and that our feeling of the same changes as they occur is the emotion”. Emotion is a feeling of a bodily state. Emotions can be released out by facial expressions, attenuation and amplification in voice, hand gesture and physiological signals. Emotion identification analysis has become a popular research area for human computer interaction over the past decades. Automatically recognition of these expressions can be achieved by the intelligent computers which are able to understand, interpret and respond to human intentions, emotions and moods [1]. Emotions can be expressed mostly by using facial expressions. “The face is like a switch on a railroad track,” says Fridlund; “it affects the trajectory of the social interaction the way the switch would affect the path of the train” [2].

Facial expressions are such communication channels that hold sense signals so they are as much as important as speech and gesture. Sometimes they are much more meaningful because they occur due to the contradictions of the muscles so reflects the natural feelings [3]. That is why it is appreciated to work on recognition of facial expressions.

1.1. Literature Survey

Facial expressions help carrying social information among humans. It plays an important role in face to face communication. Identifying facial expressions is used wherever humans interact with machines. There exist many different approaches for facial expressions recognition in the literature. In 1971, Ekman and Friesen proposed facial expressions can be clustered into six primary emotions that pose each discretely [4].

These prototype emotions are said to be basic universal emotions. In 1978 Suwa et al. proposed preliminary approach for automatic facial expressions from a sequence of images [5]. In the 1900s, Mase and Pentland caused acceleration on automatic facial expression analysis research [6]. Bourel and Chibelushi used weighted k-nearest neighbor classifier for six basic facial expressions and obtained successful recognition rate with a little degradation [7]. Facial Action Coding System (FACS) is developed by Ekman and Friesen to distinguish and code each action that can be done via face, to action units (AU) [8]. Using artificial neural network (ANN) architecture is a very suitable approach for identifying facial expressions [9], [10], [11]. A constructive neural network method was proposed on four facial expressions (angry, surprise, smile and disgust) by Ma and Khorasani with the best recognition rates of 100% and 93.75 (without rejection) [10]. Principle Component Analysis (PCA) and Gabor wavelets are also in harmonious with ANN to get better performance [12], [13], [14]. A performance 89.11% is achieved in [14]. Hidden Markov Model (HMM) approach is used and 85% success is obtained in recognition [15]. Nearest neighbor classifier and template matching methods are used and best performance is achieved by using Gabor wavelet representation 93.30% [16]. K-nearest neighbor approach is also proposed by Sayeed et al. with a success rate of 90.76% [17]. Using ANN as a classifier is a well-known fact that it provides a good performance on identification problems. Many different successful approaches are proposed under the supervision of facial expressions. Donato et al. combined PCA and optical flow [16]. They used three methods as: optical flow, PCA and high gradient component 6 upper faces FACS, AU 1, 2, 4, 5, 6 and 7 where ANN used for recognition. For 80 image sequences and 40 images of 20 objects 91% rate of success was calculated. Otsuka et al. proposed a model by combining HMM with optical flow and estimated the condition of facial muscle from image sequences [18]. By Mase major features of a face were represented by muscle regions [19]. 12 of 44 facial muscle movements were extracted by using optical flow technique. In each local region, mean and covariance of the optical low are computed. Five expressions (smile, angry, surprise, disgust and unknown) were recognized by an 86% success rate. The work done in [20] and [21] are similar to the approach of Mase, but instead of dealing with muscle groups, brows, eyes, and mouth are used as features. Optical flow technique is used to track the motion directions. This study is based on the mid-level representation. In [21] the recognition rate was 88% and 65% with eye blinking. In [20], the work done in [21] was developed for smile and surprise expressions with the addition of Radial Basis Function (RBF) and 88% recognition success was achieved. In [22] a similar approach with [21] is performed which is based on the mid-level index of motion direction of each features (mouth, eyes, and brows). By using six universal expressions for 70 image sequences of 40 subjects, 92% recognition rate is performed. Kobayashi and Hara [22], [23] and [24] used three sets of ANN for mixed combinations, 6 universal facial expressions and intensity of each expression. For 10 objects, with mixed combinations of facial expressions the success rate is 70%, for six universal expressions the recognition rate is 88.7% for 15 objects. Ding et al. used three ANN to recognize the expressions of brows, eyes and mouth [25]. Due to the symmetry fact of a face, recognition performed on the left half of the face. In references [26], [27] and [28] fuzzy logic was performed where [29] is combined with ANN to identify six basic facial expressions. Cottrell and Metcalfe applied PCA and backpropagation ANN to recognize the facial expressions and gender from static images [30]. Also, Cottrell and Padgett proposed a similar approach to recognize facial expressions by combining PCA and feed forward ANN [31]. And the hit ratio is increased to 84%. Matsuno et al. also used four expressions (smile, angry and surprise) with new method of recognizing facial expressions using a two dimensional physical model named *potential net* [32].

ANN is a suitable method that can be used to recognize facial expressions. It is also a type of artificial intelligence that approximates the operations of the human brain. ANN structure consists of interconnecting processing elements which are called units [33]. These units are the analogous of the neurons in the human brain. The output of ANN is determined by the organization and weights of the connections. To obtain the target output ANN is trained by a large number of data and rules about the relationship. So the program learns how to behave in response to an external stimulus.

Currently ANN is mostly used in image recognition, voice recognition, medical imaging, and industrial robotics. Nowadays by the information technology we can save large quantity of data. Every moment there are millions of data that reach to their central data storage. These data come from the safe of the supermarkets, the automated teller machines from which we withdraw money and the e-commerce applications. Expectation of analyzes of these applications are different. For example a manager of a supermarket wants to learn which of the products are sold together, an analysts of a exchange market wants to learn the worth of tomorrow's allotment, an iris recognition system wants to learn the person of that iris, a security camera wants to know whether there is an abnormal event or not. In all these applications previous data is used to process new data. It is impossible to process and analyze large quantity of data by hand. In order to solve this problem machine learning methods had been developed. Machine learning methods tries to find the most suitable model for the new data by using a previous data. To understand whether a signature is fake or not, to recognize a person in the image, to identify the owner of the fingerprint, to identify the owner of the voice, to figure out the hand signals that the hearing disabled use. ANN is one of machine learning approach.

1.2. Organization of the Thesis

Chapter 2 presents background information on neural network, AR database, PCA, image enhancement and color space for image processing. Chapter 3 explains design of the first proposed method for emotion identification which is based on cascaded tree PCA+ANN structure and design results. In Chapter 4, a second method is introduced as an approach that uses only hidden layer outputs of ANN to recognize the facial expressions. The last method is proposed in Chapter 5. In this method skin detection is applied to a facial image and then to extract eyes with brow and mouth from a non-skin region cropping is applied. The design and methodology is explained in Chapter 5. Finally, in Chapter 6 the concluding remarks are stated on the performance of the system and mentioned to the future works.

CHAPTER 2

THEORETICAL BACKGROUND

2.1. Artificial Neural Network

Artificial neural network (ANN) is a type of artificial intelligence of consecutive stages that approximates the operations of the human brain. Dealing with ANN requires two parts, one of which is the training and the other is the testing. In the former, training data is used to train the ANN and in the latter the testing data is employed to assess the performance of the ANN on classification. There are several training algorithms in neural network. Multilayer Perceptron (MLP) is one of these algorithms. It uses forward propagation, backpropagation and updates its weights. ANN has been proven as a successful method in pattern recognition [34]. Facial expressions can be identified by using ANN which is among the pattern recognition. ANN structure consists of interconnecting elements, called processing units that operate in parallel. These units are inspired from a biological nervous system.

2.1.1. Biological Neuron

The human brain has billions of neurons. In a human brain there are nearly 10.000.000.000 interconnected neurons that have parallel computing capability. A biological human nerve cell is shown in Figure 1.

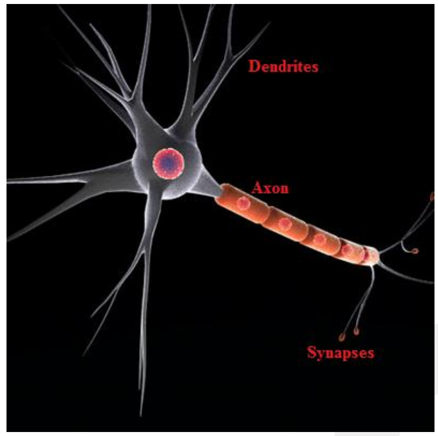


Figure 1: Biological neuron [35].

These neurons transfer stimulus coming from human sense organ to dendrites and then flows through axons [36]. Jackson et al. states that all inputs to the cell body of the neuron arrive along dendrites, which acts as outputs interconnecting neurons and axons are located on the output cells [37]. There is no direct connection between axons and dendrites. The gap between them is called synapse. In synapses the stimulus transfer is done by chemical ways; where in dendrites, this transfer is obtained by an electrical process. In excess (fire) of a certain threshold or an activation level as transferring stimulus, it generates an electrical and chemical change in neurons which is called impulse. These neurotransmitters generate impulse to transmit inputs to the next neuron, so provides the message transfer between neurons. Receiving of signals between neurons is simply done by this way. As a result of the collective effects of this process thinking, learning, perception and cognition occur.

2.1.2. Neural Network Neuron

In human brain there exist billions of neurons. ANN is an artificial model of biological neuron which is inspired from human neuron. It consists of multiple inputs with a single output. To better understand how ANN works, let's examine the difference between human brain and modern computer. In modern computer there is highly complex processing unit with high rate of process speed. In human brain, which has already been mentioned above, there exists billions of processing units (neurons) which are simpler and slower in process compared to modern computer. To denote the fire of stimulus in a human brain with a certain threshold value binary logic is computed in computers. In binary logic, activation is denoted by “1” and inactivation is denoted by “0”. In ANN when a neuron fires, then it passes up information that it has already received through the network chain. Modeling the neurotransmitter transfer between neurons can be obtained by modeling neurons as switches. The synapses between human neurons can be corresponds to the numbers (weights) of matrices in ANN.

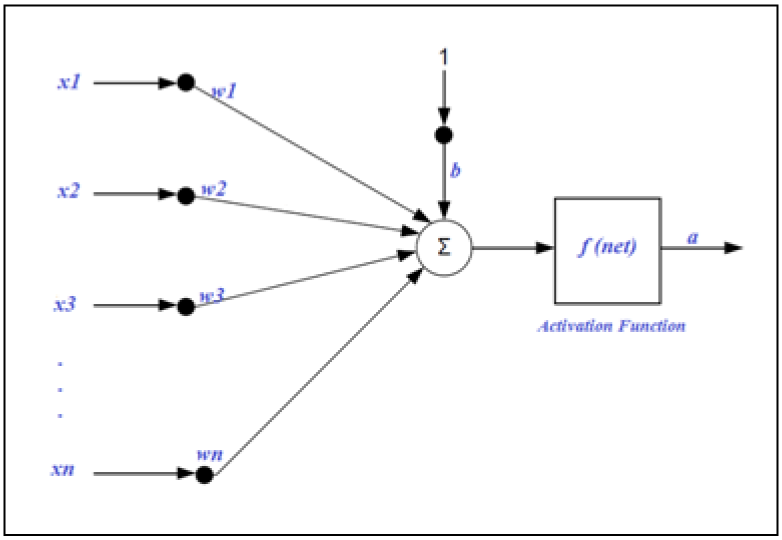


Figure 2: ANN with activation function.

As it is seen in Figure 2, inputs are modified by updating the corresponding weight values and then combined by a neuron. Output is determined by neuron with the reference of threshold value or activation function. This process runs in a similar manner with biological neurons and the similar functions due to this process are generated.

2.1.3. Learning

Although it is not possible to model a human brain exactly with its full complexity, an ANN can be used to solve high complexity problems of human brain. This can be obtained by training ANN architecture. By fully training the ANN learning is achieved. The learning mechanism of ANN is similar to the speaking mechanism of a new born baby. After a born, baby starts to records the words that (s)he heard. It is an iterative and continuous process until the baby successes spelling words correctly. This process corresponds to the learning mechanism of neural network terminology.

2.1.3.1. Supervised Learning

In supervised learning outcomes for known inputs are tried to be predicted. The most common type is backpropagation (BP). Supervised learning is learning from mistakes. System compares its predictions with the target output, and converges to the target by learning from its mistakes. Weighting is applied through the inputs then inputs pass through the next node, in there the updated weights are summed up and either intensified or weakened. Initially the system starts with random weights, makes an initial guess about the correct output, get the difference, known as error, between the guess value and the correct output. To minimize the error rate the weights are changed and this iterative process of prediction continues until the data reaches to the output layer. In here it is compared with the actual output and if they are equal no change is applied on weights, else error is propagated through the system and weights are updated until the target is obtained or nearly approximated. Feeding the system with the error is known as BP where it will be detailed in 2.1.7.

2.1.3.2. Unsupervised Learning

Unsupervised learning differs from supervised learning in describing data rather than predicting. In this learning method NN does not have any idea about the actual output. There is no output field in this architecture. Kohonen network (self-organizing map) is the mostly known type of unsupervised learning. Unsupervised learning is an ideal process for clustering similar data. Simply, supervised learning is based on training data where unsupervised learning is learning without training data. Based on the type of the problem these two methods can also be mixed up to obtain the best performance.

2.1.4. Activation Function

Activation function defines the output of a node by transforming the action level of it. There are four main types of activation function as threshold, linear, logistic sigmoid and tangent sigmoid.

2.1.4.1. Threshold Activation Function (McCulloh-Pitts Model)

Threshold activation function is the early model of ANN which was introduced by McCullon and Pits in 1943. It is a kind of a linear threshold gate. The output is hard limited to either 1 or -1 (or 0) due to the sign of the net. Matlab generated sample threshold function is given in Figure 3.

$$n = \begin{cases} +1 & for\ a > 0 \\ -1 & for\ a < 0 \end{cases} \tag{2-1}$$

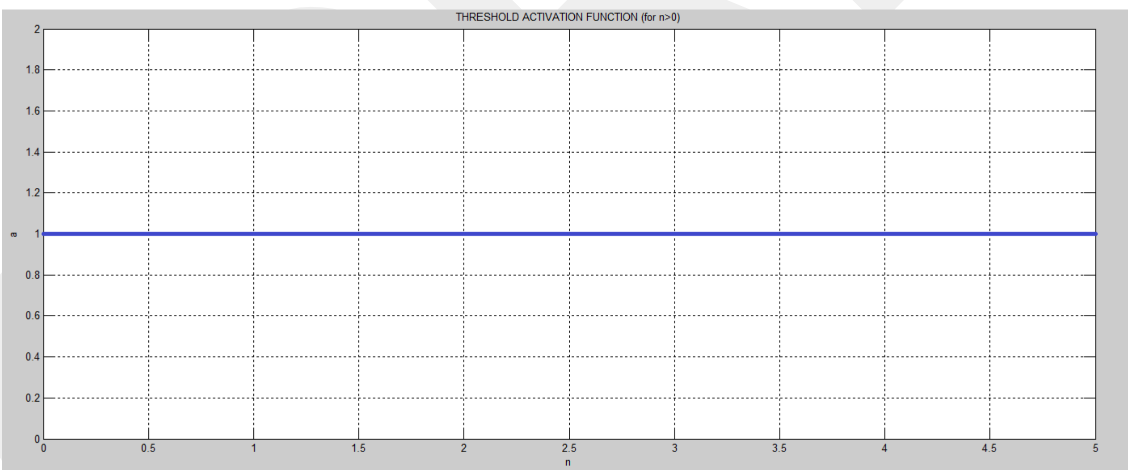


Figure 3: Threshold activation function.

2.1.4.2. Linear Activation Function (Purelin)

Depending on the input, output changes linearly as it is seen in Figure 4. The dynamic change range of this function is from -1 to 1.

$$a = f(n) \tag{2-2}$$

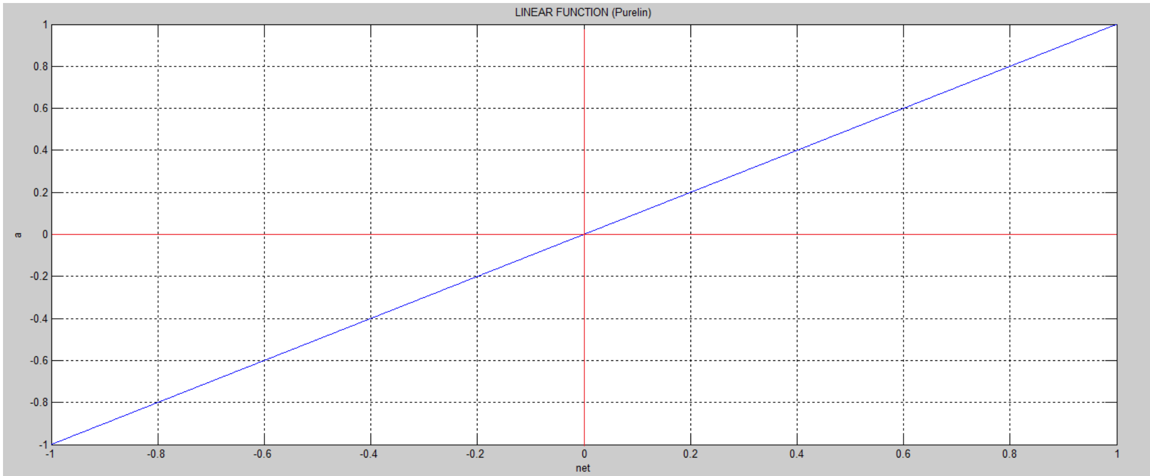


Figure 4: Linear activation function.

2.1.4.3. Logistic Sigmoid Activation Function

Sigmoid function which is also referred to ad logistic sigmoid function is the widely used type of an activation function of ANN. Despite McCullon and Pitts model, sigmoid function assumes a continuous range of values from 0 to 1 with a nonlinear change as it is shown in Figure 5. It is a differentiable function which exhibits an important property of ANN theory.

$$a = \frac{1}{1+e^{-n}} \tag{2-3}$$

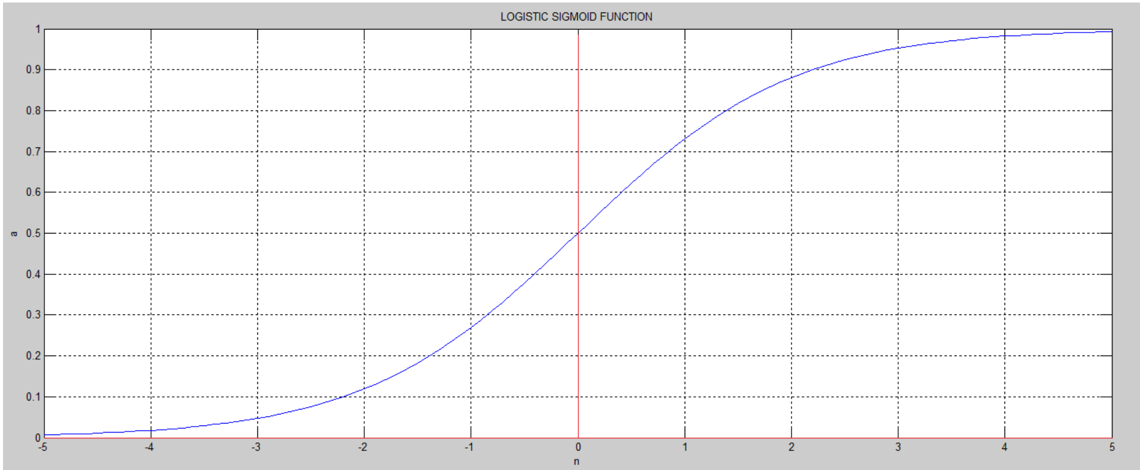


Figure 5: Logistic sigmoid activation function.

2.1.4.4. Tangent Sigmoid Activation Function

As it is shown in Figure 6, the dynamic change range of this function differs from -1 to 1 and depending on the input data it exhibits a nonlinear change in this range.

$$a = \frac{1-e^{-n}}{1+e^{-n}} \tag{2-4}$$

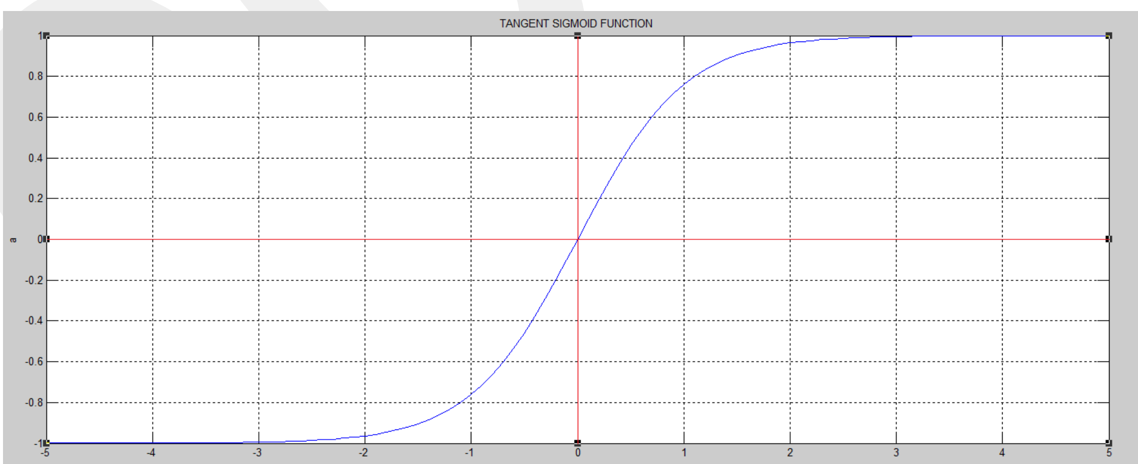


Figure 6: Tangent sigmoid activation function.

The hyperbolic tangent function simply equals one less of two times of logistic sigmoid function as it is stated in Equation (2-5).

$$2\left(\frac{1}{1+e^{-n}}\right)-1 \tag{2-5}$$

2.1.5. Perceptron

It is the basic form of ANN which is used to classify linearly separable patterns. Perceptron utilizes supervised learning algorithm that responds either 1 (true) or 0 (false). Structure of perceptron, which is given in Figure 7, is simple with a McCullon and Pits neuron, adjustable synaptic weights and a bias value.

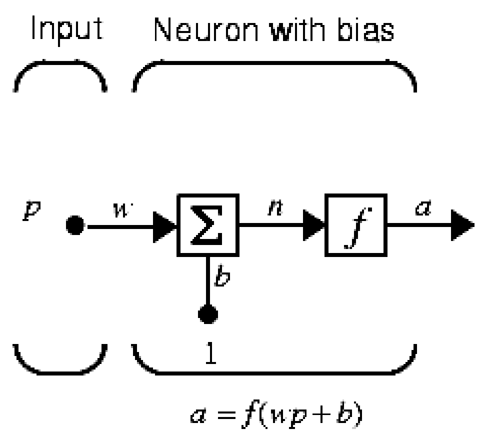


Figure 7: Single neuron architecture [38].

In this architecture, bias and each input values are connected to the main perceptron by a weight. Inputs are multiplied by corresponding weights, then summed up and passed through the hard limiting function, which is simply a threshold setter to fire the perceptron. Also it is called a Threshold Logic Unit (TLU). It is capable of separating input spaces into two hyper planes. Consider AND gate. It is obvious that it is a linearly separable problem. For two linearly separable classes the decision boundary is drawn in Figure 8, so that the classes can be separately classified as A and B.

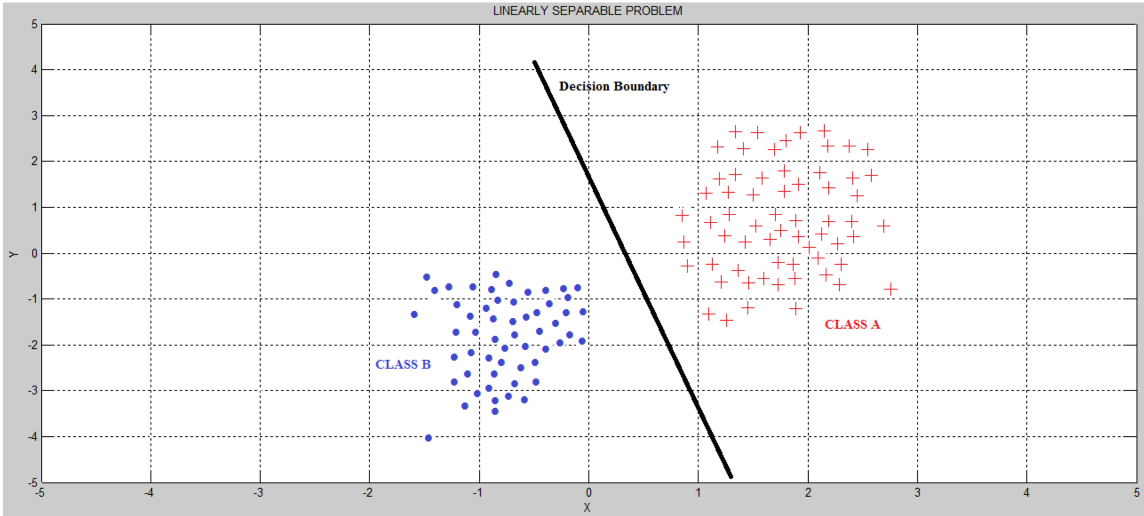


Figure 8: Linearly separable problem.

In perceptron algorithm the basic learning is performed by adjusting weight values by the difference between target output and the actual output. In this algorithm learning is guaranteed by the perceptron convergence fact that is stated by Rosenblatt [39].

2.1.6. Multilayer Perceptron (MLP)

Perceptron is capable of solving only linearly separable problems. On the other hand there exist many more challenging problems which are not linearly separable in the real world. Perceptron has a weakness to solve these kinds of complex problems. Consider XOR problem. Single layer does not sufficient to separate the classes. Minsky and Papert offered a solution to XOR problems [40]. The idea is to convert nonlinear problem to a linear problem by combining the perceptron unit responses using a second layer of units [41]. In Figure 9, two nonlinearly separable data cluster are given, since they are nonlinear no exact decision boundary can be drawn to separate them into two classes.

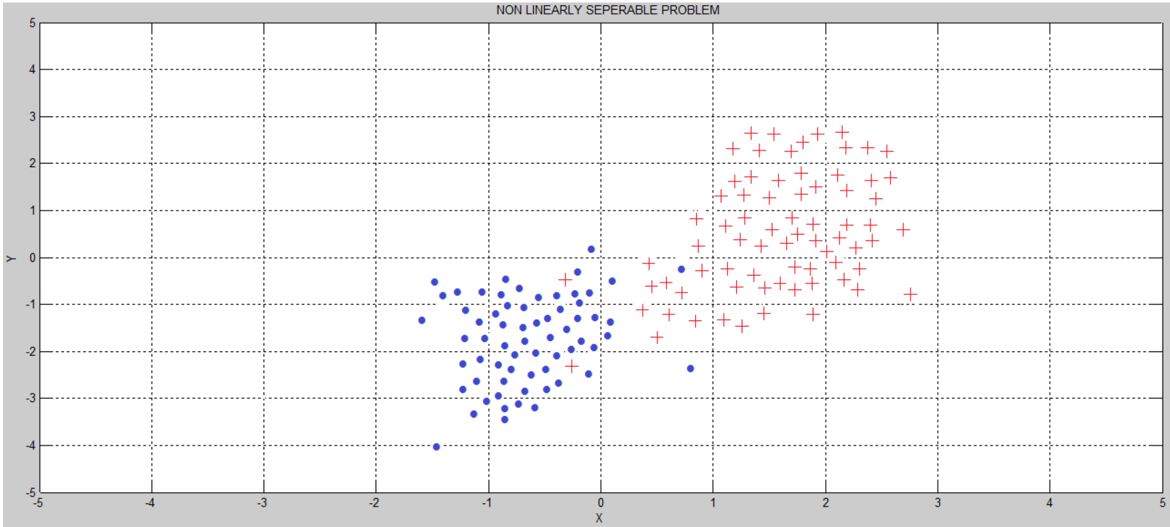


Figure 9: Nonlinearly separable problem.

MLP ANN composed of units that are arranged in layers [42]. In that arrangement each layers consist of minimum three layers as: input, hidden and output layer. Input data is distributed through the network via input layer. Three-layer MLP can be able to approximate any continuous function as stated in [43] and [44]. Hidden layer nodes have nonlinear activation function and thresholds are used. In this cascaded architecture, original input, multiplied by weight with a threshold added, flows through an activation function.

2.1.7. Backpropagation (BP)

It is a supervised learning algorithm that is based on delta rule by minimizing the error between the actual output of the ANN and the target output. The main principle of this algorithm is to arrange weights to provide outputs which are obtained by special function characteristics that correspond to the input data. The error is calculated by the difference between ANN output and the target output. If the error is large, then it is fed back to the ANN to update synaptic weights in order to minimize the error. This process continues until the minimum error is reached. BP has two phases as propagation and weight update. Propagation phase has two steps as forward and backward pass as it is shown in Figure 10. Forward pass feed forwards the input pattern signal through the network by computing functional signal [41]. By starting from the output layer backward pass propagates the error backward which is the difference value between actual and target. This process continues until the minimum error is reached. In weight update phase, input activation level and output delta are multiplied to get the gradient weight. Then weights are put in the reverse direction of the gradient by subtracting the ratio of it from the weight.

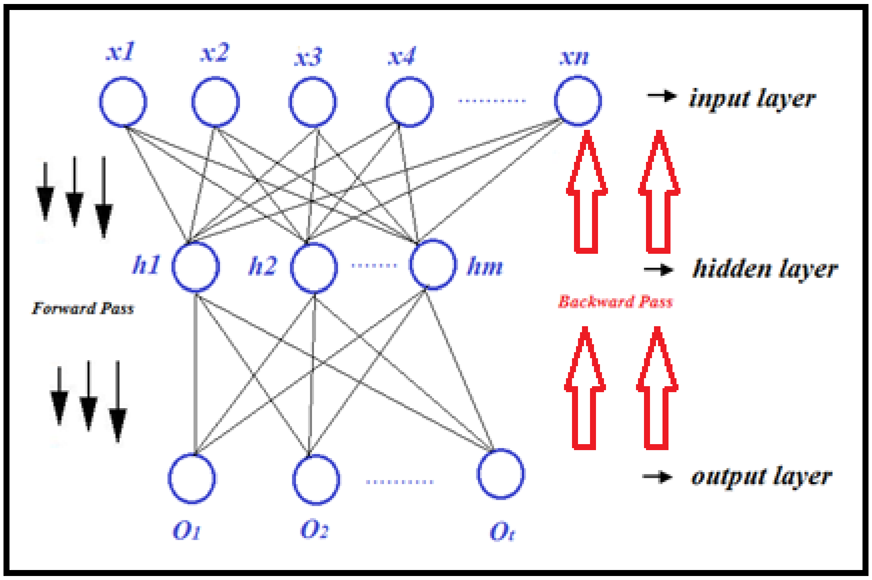


Figure 10: MLP with backpropagation.

The network computes the error E by having the sum of the square of the difference between the activity level and the target output.

$$E = \frac{1}{2}\sum_{j}(x_j - t_j)^2 \quad (2\text{-}6)$$

x_j : is the activity level of the j^{th} units in the output layer.

t_j : is the target output of the j^{th} unit. The total weighted input of the net is equal to net_j

$$net_j = \sum_{i} x_i w_{ij} \quad (2\text{-}7)$$

where w_{ij} is defined as a weight value from unit *i* to unit *j*. Calculation of activity level is hold by using sigmoid function.

$$x_j = \frac{1}{1+e^{net_j}} \tag{2-8}$$

2.1.7.1. Backpropagation Algorithm

The algorithm can be summarized in four steps.

1. Computation of the change rate of the error function as the activity level changes by calculating derivative of the error function EA_j (difference between the actual and the target output).

$$EA_j = \frac{\partial E}{\partial x_j} = x_j - t_j \tag{2-9}$$

2. Computation of the change rate of the error function as the input entered through the output unit by calculating EI_j.

$$EI_j = \frac{\partial E}{\partial net_j} = \frac{\partial E}{\partial x_j}\frac{dx_j}{dnet_j} \tag{2-10}$$

$$= EA_j x_j (1-x_j) \tag{2-11}$$

where net_j is the total weighted input

$$net_j = \sum_i x_i w_{ij} \tag{2-12}$$

with x_i activity level of the i^{th} unit of the previous layer and w_{ij} is the weight of the connection between i^{th} and j^{th} units where x_j is the activity of the j^{th} node.

$$x_j = \frac{1}{1+e^{net_j}} \tag{2-13}$$

3. E is minimized by using an iterative process of gradient descent.

$$\nabla E = \left(\frac{\partial E}{\partial w_{11}}, \frac{\partial E}{\partial w_{12}}, \frac{\partial E}{\partial w_{13}}, ..., \frac{\partial E}{\partial w_{1j}} \right) \tag{2-14}$$

Each weight is updated by using the increment of

$$\nabla w_{ij} = -\gamma \frac{\partial E}{\partial w_{ij}} \tag{2-15}$$

Simply the weight update is performed as following $w_{new} = w_{old} - \nabla w_{ij}$ where γ is the learning constant that defines the step parameter of step length.

Computation of the EW_{ij} which is the change rate of the error function as a weight on the connection through the output unit

$$EW_{ij} = \frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial net_j} \frac{\partial net_j}{\partial w_{ij}} \tag{2-16}$$

$$= EI_j x_i \tag{2-17}$$

4. These 3 steps, iteratively goes on until the error is minimized

$$EI_i = \frac{\partial E}{\partial x_i} = \sum_j \frac{\partial E}{\partial net_j} \frac{dnet_j}{dx_i} \quad (2\text{-}18)$$

$$= \sum_j EI_j w_{ij} \quad (2\text{-}19)$$

2.2. Dataset

In literature, there have been commonly used two dimensional (2D) databases such as FERET [45], CMU-PIE [46], JAFFE [14], [17], [47], [48], AR [49], [50] and the others [51]. In this thesis, AR database is used [49]. There are 126 people's face images (70 men and 56 women). Images are captured from distinct people’s faces with 4 facial expressions of each. All are frontal view faces with different facial expressions, illumination conditions, and occlusions (glasses, mustache, beard, fringe etc.). The images are totally benchmark. The expressions of neutral (C_1), smile (C_2), angry (C_3) and scream (C_4) are used from AR database. The images are in red, green and blue (RGB) format. Each color is defined by a three dimensional (3D) color space.

In this thesis study some color space conversations are applied to the original RGB images. For the first two studies RGB images are converted into gray scale (GS) images, where in the third one YUV and HSV color space conversations are applied. In Figure 11, GS converted images from AR database are given.

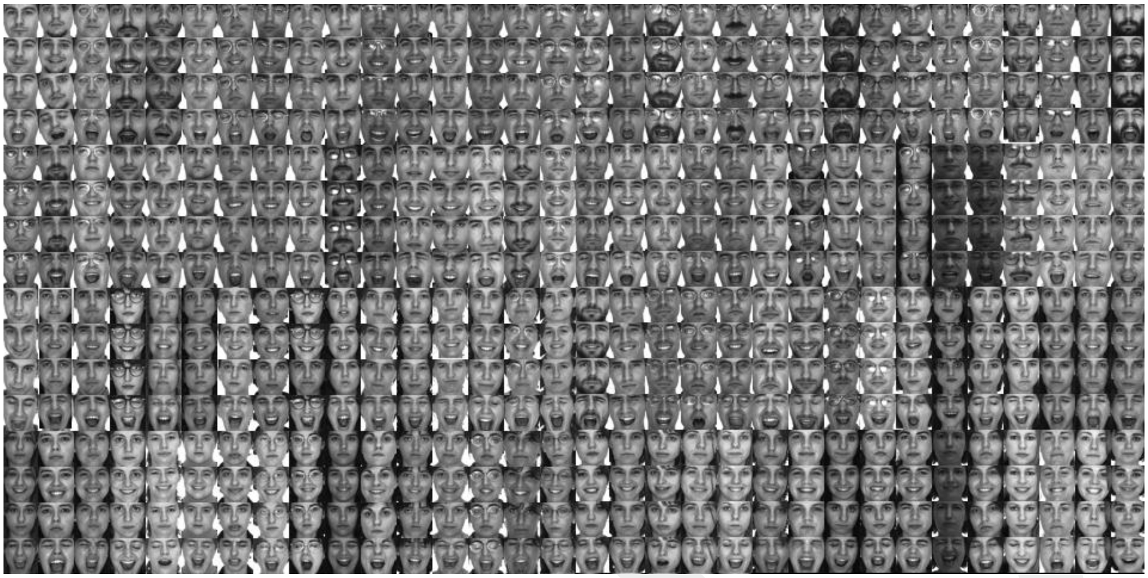


Figure 11: GS images from AR database.

The original images are in size of 768x576 are cropped into the size of 250x250 by ensuring that all eyes are aligned. Then cropped images are resized into 50x50 images.

2.3. Principle Component Analysis (PCA)

PCA is a local image based approach that is commonly known as subspace method for appearance based recognition [52]. It is invented by Pearson and Hotelling. It is based on the correlation between the image pixels. PCA converts possible correlated observations into a set of values of linearly uncorrelated variables by using orthogonal transformation. These linearly uncorrelated variables are called as principal components. PCA is so-called as Karhunen-Loève transforms which is a transformation to express data in lower dimensions with minimum error in the mean square error sense [53]. By finding the orthogonal linear combinations of the variables with the maximum variance it reduces the dimensions. In the case of emotion identification by using facial expressions, PCA can be efficiently used. Instead of studying on a whole image, using only the difference between individual images is enough for identification. The mostly used method for emotion identification is eigenface method which is based on PCA.

2.3.1. Eigenface Approach

Eigenface is a ghostly face of a real image. This approach takes the whole image and encodes it to a value. The idea is representation of an image with its significant features, which include the highest eigenvalue inside. Any image can be almost reconstructed by a small collection of its weights. Projection of the face image onto the subspace of eigenfaces generates these weights. By comparing these weights with the corresponding weights of each facial expression, recognition can be done. Recognition of a test data is obtained by projecting it on the eigenfaces subspace and then compares its weights with the corresponding weights of each image class from the database. Calculating principle components of an image distribution means calculating the eigenvectors of the covariance matrix of the set of facial expressions classes. Eigen is a word that is derived from the German and stands for *characteristic*. So, simply eigenvector is a set of vector that holds the characteristics (features) of an image. For each eigenvector there exists a corresponding eigenvalue. This eigenvalue is the factor by which the eigenvector is multiplied when transformed. They are in an ordered fashion with different amount of variation among the faces. According to the need of the study, the most meaningful eigenfaces are chosen. The threshold for this choice is the saturation. Because after a while eigenfaces reach a saturation. The eigenfaces in this saturated area are redundant to use. In Figure 12, the top three eigenfaces from AR database are given.

Figure 12: First three eigenfaces of AR dataset.

Consider sample of n observations on a variable of p variables $\{x_1, x_2, \cdots, x_n\} \in \Re^p$. The centroid (mean) of the points is of the form;

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2-20}$$

$x_0 = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean that represents the data set by a single point x_0. To justify this choice, we need to find the solution that minimizes J_0.

$$J_0(x_0) = \sum_{i=1}^{N} \|x_i - x_0\|^2 \tag{2-21}$$

Data can also be represented by a line of the form that passes through a point of m.

$$x = m + ae \tag{2-22}$$

where e is the unit vector along the straight line and a is the signed distance of the point x from m. The training data projected on this line is shown by;

$$x_i = m + a_i e, \quad i = 1...n \tag{2-23}$$

by partially differentiating with respect to a_i

$$J_1(a_1, a_2, ..., a_n, e) = \sum_{i=1}^{n} \|m + a_i e - x_i\|^2 \tag{2-24}$$

$$= \sum_{i=1}^{n} a_i^2 \|e\|^2 - 2\sum_{i=1}^{n} a_i e^T (x_i - m) + \sum_{i=1}^{n} \|x_i - m\|^2 \tag{2-25}$$

$$= \sum_{i=1}^{n} a_i^2 - 2\sum_{i=1}^{n} a_i e^T (x_i - m) + \sum_{i=1}^{n} \|x_i - m\|^2 \tag{2-26}$$

and a_i is obtained as

$$a_i = e^T(x_i - m) \quad (2\text{-}27)$$

Substitute a_i in J_1

$$J_1(e) = -\sum_{i=1}^{n} e^T(x_i - m)(x_i - m)^T e + \sum_{i=1}^{n} ||x_i - m||^2 \quad (2\text{-}28)$$

$$= -e^T \sum_{i=1}^{n} (x_i - m)(x_i - m)^T e + \sum_{i=1}^{n} ||x_i - m||^2 \quad (2\text{-}29)$$

where $\sum_{i=1}^{n}(x_i - m)(x_i - m)^T$ is a semi defined matrix which is scatter matrix S which is also $n-1$ times of the covariance matrix.

2.3.2. Evaluating Eigenvalues

Minimizing J_1 results in maximizing $e^T Se$ depending on the constraint that e is the unit vector, $e^T e = 1$. Evaluating Lagrange multiplier method by differentiating with respect to e, the eigenvalues are calculated

$$e^T Se - \lambda(e^T e - 1) \quad (2\text{-}30)$$

$$Se = \lambda e \quad (2\text{-}31)$$

where e is the eigenvector of the S that corresponds to the largest eigenvalue of λ. An eigenspace of S is the set of all eigenvectors with the same eigenvalue with the zero vector where this zero vector is not an eigenvector.

2.4. Image Enhancement

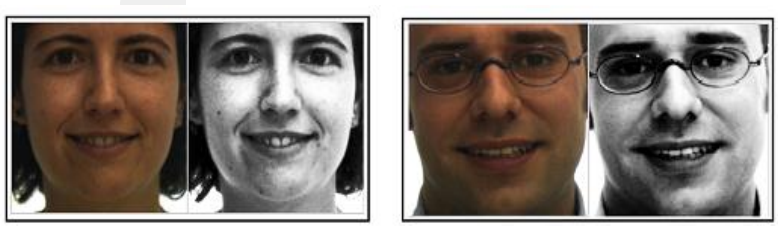
Image enhancement is the process of adjusting digital images so that the results are more suitable than the original image for display or further analysis [54]. It provides easier identification of key features by eliminating noise or brightening the image. For example, by image enhancement you can remove noise or brighten the image, so that the key features can be identified easier.

2.4.1. Histogram Equalization

Histogram equalization (HE) is a preprocessing technique that enhances contrast in natural images. Consider an image, for every gray value it consists of an equal number of pixels, but histogram is constant. For an image of size M by N with intensity range [0, L-1], then the normalized histogram is,

$$p(r_l) = \frac{n_l}{NM} \qquad l = 0,1,2,\ldots,L-1 \tag{2-32}$$

where r_l is the l^{th} intensity value and n_l is the number of pixels with the intensity of r_l. Basically, HE gives information about the statistics of an image where $p(r_l)$ is the estimate of the probability of occurrence of intensity level r_l in an image [54]. In Figure 13, HE is applied to two sample images from AR dataset. As it is seen from the figure HE provides enhanced image by removing noise and by brightening the image, so that the key features of a face can be able to recognized easier.



(a) **(b)**

Figure 13: RGB vs HE images.

2.5. Color Space Conversion

The reflection of a light from an object is perceived as color. Robertson studied on color perception and discussed color and human visual system relationship [55]. Color is the reaction of a brain to a specific visual stimulus. Color space is a method which human can create, specify and visualize color. In Figure 14, color spectrum of white light is shown. As it is seen from figure white color is a nano-metric value which is composed from different color wavelengths as; red, orange, yellow, green, blue and violet.

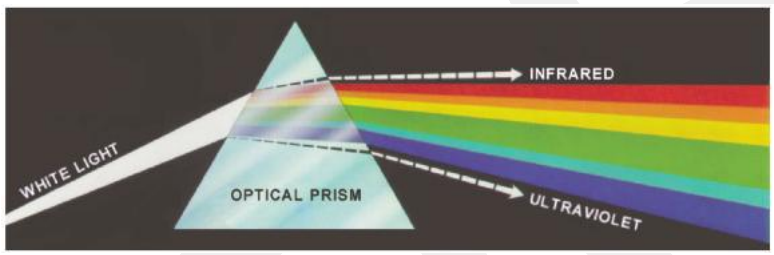


Figure 14: Color spectrum of white light passing through prism [54].

Humans recognize color due to its brightness, colorfulness and hue, where computer recognize color due to red, green and blue phosphor emissions. There are more than one color spaces in literature because different color spaces are suitable for different applications.

2.5.1. RGB (Red, Green and Blue)

RGB format is composed of red, green and blue; so it has three dimensions. This format is mainly used in system that is based on cathode ray tube (CRT), such as televisions (TV), in order to display image. It is the most common color space ranges in between TV to computer, video to projectors, etc.

2.5.2. HSV (Hue, Saturation and Value)

HSV color space defines colors as hue, saturation and value respectively. Hue stands for the dominant wavelength of the color such as red, green, purple and yellow. Saturation is the brightness of a color and value is the rate of whiteness in a color. In 1978, HSV color space is introduced by Ray Smith. HSV is the most similar color space to the human perception. To generate high quality images mostly HSV is preferred. It is used to select various different colors which are needed for an image. Hue and saturation hold the necessary information about the skin color. The skin color pixel should satisfy the following condition [56].

$$0 <= \textbf{\textit{hue}} <= 0.25 \land 0.15 <= \textbf{\textit{saturation}} <= 0.9 \tag{2-33}$$

2.5.3. YUV (Luma and Chrominance)

YUV color space encodes a color image by taking human perception into account. It allows reducing bandwidth of chrominance components. YUV format provides efficient masking to transmission errors and compression artifacts. Y' stands for luma and UV stands for two chrominance components. Luma represents the brightness of an image. UV signals contains color information in chrominance which is a signal generally used in video systems to convey the color information a picture. It is represented by two color difference components as blue difference (***C_b***) and red difference (***C_r***). For skin detection the skin portion of an image should satisfy as follows [56].

$$140 <= \boldsymbol{C_r} <= 165 \land 140 <= \boldsymbol{C_b} <= 195 \tag{2-34}$$

$$140 <= \textbf{\textit{R'(red)-Y'(luma)}} <= 165 \land 140 <= \textbf{\textit{B'(blue)-Y'(luma)}} <= 195 \tag{2-35}$$

CHAPTER 3

TREE BASED NEURAL NETWORK DESIGN

3.1. Introduction

In this study tree based neural network design for facial expressions identification is proposed [57]. AR database is used to train and test the neural network. The images are in RGB format originally. For this study they are all converted into grayscale. As it is mentioned in Part 2.2 they cropped into 250x250 images. The cropped images are resized in 1/5 scale and prepared as inputs. Then these input matrices are processed by mapping each row's means to 0 and deviations to 1. Three scenarios are presented in this study. In the first one, one unit PCA+ ANN structure is used and the confusion between neutral and angry expression is defined regarding the confusion matrices. In scenario 2, due to this confusion, neutral expression (which can be thought as a basis of each expression) is eliminated then the same procedure is applied and better recognition rate is obtained. In the Scenario 3, a new approach is proposed as tree based neural network design. By this new approach, despite of using four expressions, confusion is overwhelmed by using cascaded tree PCA+ANN structures.

3.2. Scenario 1: One Unit PCA+ANN Structure

In this study ANN architecture is fed by PCA. As it is mentioned in 2.3, PCA uses an orthogonal transformation to convert observation data set of possibly correlated variables into a set of linearly uncorrelated variables which are called principle components. PCA is used to reduce the dimension of the data set which is inputted through an ANN. 94 meaningful eigenfaces are chosen as an output of PCA as it is shown in Figure 15.

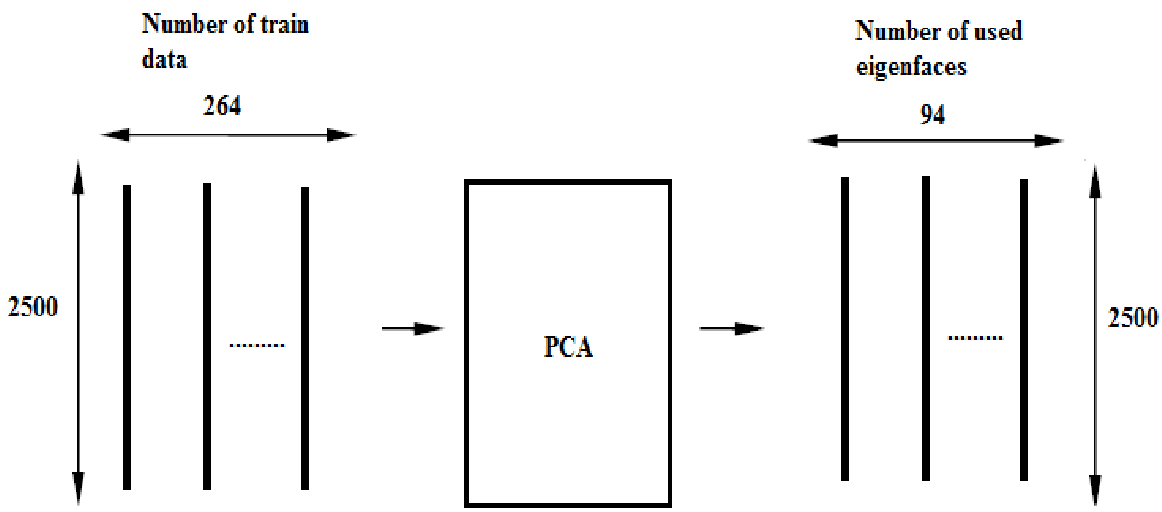


Figure 15: Generation of eigenfaces.

In this study whole image is used as an input to PCA+ANN, so PCA helps to reduce the dimension of the input into a smaller size by eliminating redundant data inside. For 66 distinct people with four expressions of each (angry, neutral, smile and scream), there are totally 264 images in a dimensions of 50x50. PCA is used to reduce input size of 2500x264 into lower dimensions of 2500x92 which also reduces the duration of the process. Basically *processpca* function is used for PCA. Eliminating redundant data by using PCA provides extraction of key features of an image, so the ANN learning process becomes easier. And also, the size of the input vectors may be reduced by retaining only those components which subscribed more than a specified fraction (which is *min_frac* in Matlab) of the total variation in the data set. For both in this Scenario and in the rest *min_frac* value is used as 0.001. In Figure 16, there is a sample of eigenface (average image) for smile expression set.

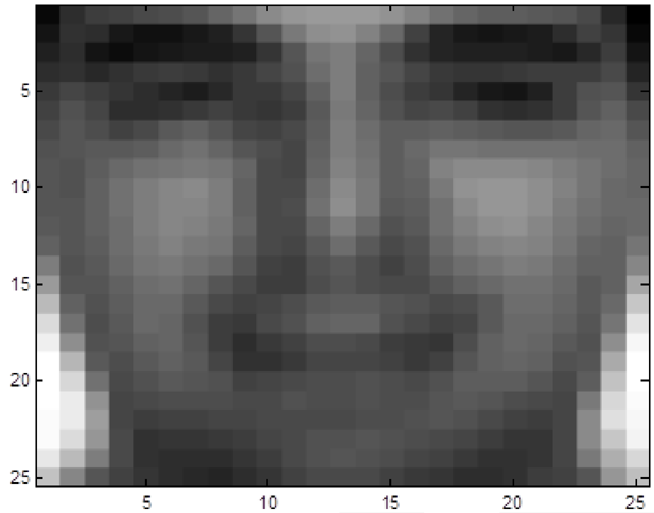


Figure 16: Average face for smile class.

In this architecture the output of PCA is used as an input of ANN with one hidden layer. BP algorithm is used to train the neural network. Inside ANN, logistic sigmoid functions are used for activation in hidden layer and *purelin* function in output layer. For 276 and 764 test data, ANN is tested with different number of epochs and hidden nodes. For both cases the best result is obtained by the combination of 20 hidden units and 150 epochs. Figure 17 shows one unit PCA+ANN architecture that is designed for facial expressions identification by using facial expressions and in Figure 18, there is a zoomed view of artificial ANN box.

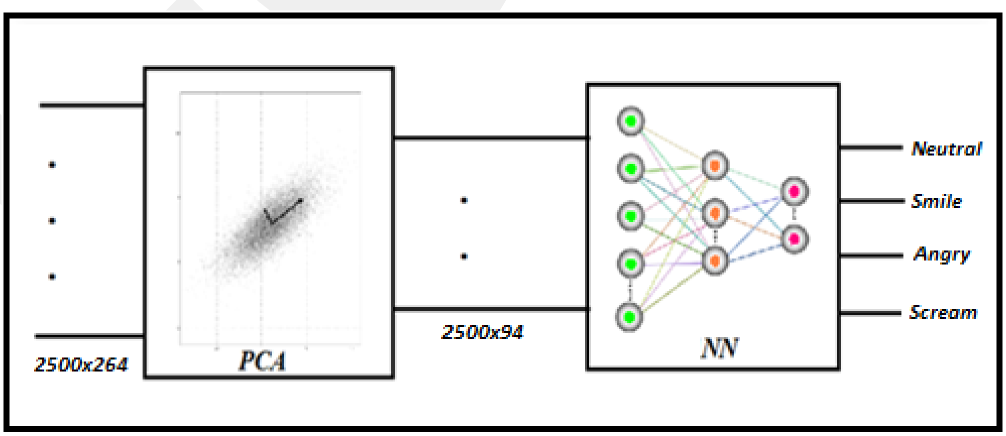


Figure 17: One unit PCA+ANN structure.

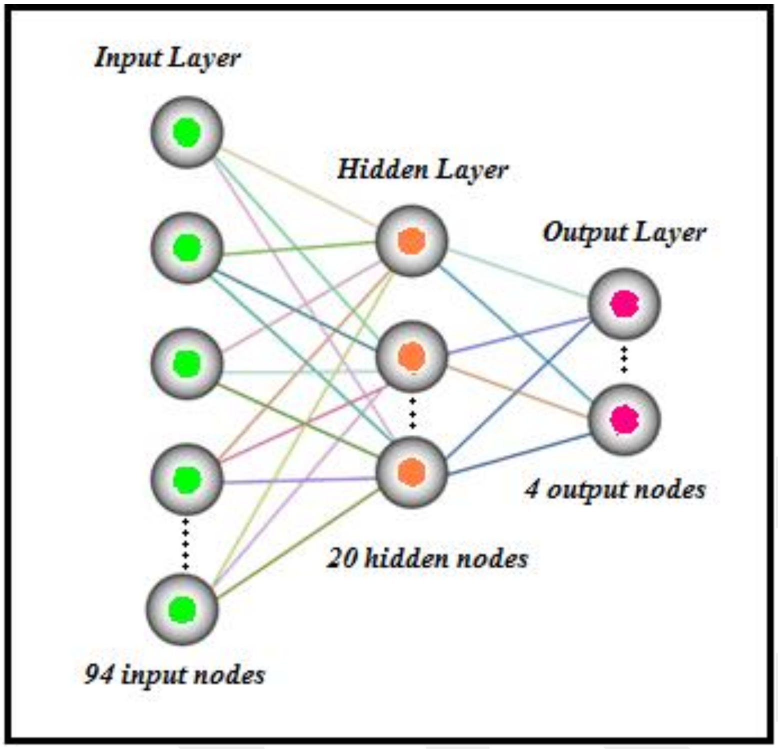


Figure 18: Inside the ANN block.

This architecture is tested with two different size of dataset. In the first test, 276 test images are used and 77.89% success is obtained. In the second test, 264 test images are used and nearly same recognition rate is obtained as 78.41%.

Table 1: Confusion matrix1 success =77.89%, miss classification=22.10%

276 Test Data	Angry	Neutral	Scream	Smile
Angry	**71.01%**	26.08%	1.54%	1.45%
Neutral	31.88%	**59.42%**	1.45%	7.25%
Scream	4.34%	1.45%	**92.78%**	1.45%
Smile	4.35%	1.45%	5.79%	**88.40%**

Table 2: Confusion matrix2 success =78.41%, miss classification=21.59%

264 Test Data	Angry	Neutral	Scream	Smile
Angry	**68.20%**	28.78%	0.00%	3.03%
Neutral	33.33%	**63.63%**	1.51%	1.51%
Scream	1.51%	0.00%	**98.48%**	0.00%
Smile	7.57%	4.54%	4.54%	**88.33%**

Including this confusion, the overall success is approximately 77% for both testing sets as they are recorded in Table 1 and 2. The results of the confusion matrix for one unit PCA+ANN structure shows that facial expressions can be divided into two classes inside. Since it is obvious, the confusion rate of neutral and angry is meaningfully high. To overcome that confusion two another approaches are proposed in Scenario 2 and Scenario 3 to increase the performance of the recognition.

3.3. Scenario 2: One Unit PCA+ANN Structure without Neutral

As a result of confusion between angry and neutral images in this scenario neutral images eliminated from both train and test sets. Because neutral expression can be thought as a basis, that each expressions can be formed from it. Due to the empirical observation for this Scenario, it is decided to use 100 epochs and 20 hidden nodes in the hidden layer. The same architecture of Scenario1 is applied for three expressions (smile, angry and scream). Again logistic sigmoid function is used as an exciter in hidden layer.

Table 3: Confusion matrix3 success =90.34%, miss classification=9.66%

264 Test Data	Angry	Scream	Smile
Angry	**95.65%**	1.40%	2.90%
Scream	1.40%	**91.30%**	4.34%
Smile	8.70%	7.25%	**84.06%**

By eliminating neutral expression that causes confusion for recognition, the overall performance is increased by 13.34 % as it is seen Table 3. But what about using four expressions by advancing the technique that is used in these two scenarios and achieving more successful recognition rate? The answer is in Scenario 3.

3.4. Scenario 3: Cascaded PCA+ANN Structure

In Scenario 1 according to the confusion matrices, choosing images of neutral and angry as one class, smile and scream in another class can be a logical clustering. Despite of using a single ANN architecture, with three cascaded ANN the performance can be increased by classifying neutral (C_1) - angry (C_3) into a class and smile (C_2) - scream (C_4) into another. The proposed architecture is given in Figure 19.

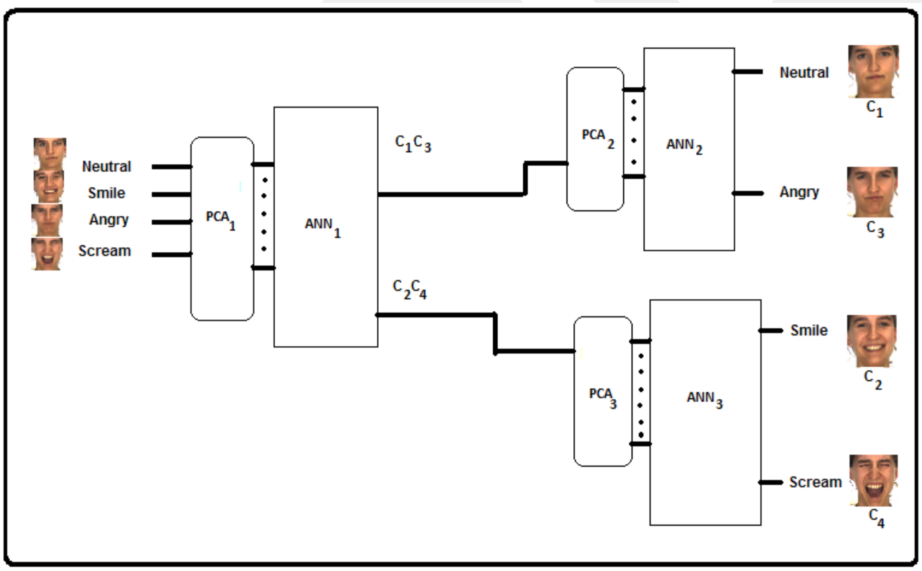


Figure 19: Cascaded tree network architecture.

In this architecture *min_fac* value is used as 0.001 for each PCA. For ANNs hidden layers logistic sigmoid functions are used as activation function. ANN_1 is trained by 100 epochs where ANN_2 and ANN_3 are trained by 150 epochs. Optimal hidden node values differ in each ANNs. According to this architecture ANN_1 is trained with 12 hidden nodes to obtain the outputs of C_1C_3 and C_2C_4. To classify C_1C_3 and C_2C_4 into C_1, C_2, C_3 and C_4, ANN_2 and ANN_3 are separately trained by 150 and 120 hidden nodes respectively. In this manner, C_1 and C_3 classes are recognized by ANN_2 where C_2 and C_4 classes are recognized by ANN_3. In Table 4, confusion between C_1C_3 and C_2C_4 is given. In Table 5 and 6 confusion matrices for each class (C_1, C_2, C_3 and C_4) are separately calculated. By this proposed tree architecture of ANN, despite the fact of confusion, recognition rate is obtained as 88.25% which is 11.25% more than the result of the Scenario 1.

Table 4: Confusion matrix4 success =98.99%, miss classification=1.005%

276 Test Data	C_1C_3	C_2C_4
C_1C_3	**98.97%**	1.03%
C_2C_4	0.98%	**99.02%**

Table 5: Confusion matrix5 success =80.88%, miss classification=19.12%

276 Test Data	C_1	C_3
C_1	**88.51%**	11.49%
C_3	26.75%	**73.25%**

Table 6: Confusion matrix6 success =95.62%, miss classification=4.625%

276 Test Data	C_2	C_4
C_2	**96.50%**	3.50%
C_4	5.75%	**94.75%**

In paper [58] whole pixel image is applied through Linear Discriminant Analysis (LDA), in this Scenario MLP with tree neural network structure is used for the same AR database and 9.05% better result is achieved. In paper [59] cascaded classification method is used for support vector machines (SVM) to classify gender and facial expressions.

3.5. Conclusion

In this study, automatically identification of facial expression is obtained by using cascaded tree ANN structure. The idea is classification of most confused expressions in a class and the rest in another class. In AR dataset, angry and neutral expressions are confused even by human perception. This confusion is overwhelmed by this cascaded tree ANN structure. This method helps to increase the training performance. Due to the success of training set, the test set performance is also increased and recognition of four expressions can be achieved with a successful rate. Although human perception is not 100% accuracy at the testing set, achieving 88.30% success should not be underestimated.

CHAPTER 4

EMOTION CLASSIFICATION USING HIDDEN LAYER OUTPUTS

4.1. Introduction

Learning types of ANN and how this learning process occurs were discussed in Chapter 2. ANN teaches a certain pattern during back propagation phase and learns when the network is fully trained. In BP algorithm error correction is the basis of achieving recognition. By providing the smallest error rate, the difference between the actual output and the target output becomes minimum, so that the algorithm converges and the recognition is performed. To achieve training many number of epochs also must be processed.

Using ANN as a classifier is a well-known fact that it provides a good performance on identification problems. On the other hand, ANN needs to be fully trained in order to solve a classification problem so the training phase may take long time due to the size of the epoch. In Mikalov et al. it is stated that the computational complexity of an ANN depends also on epoch numbers [60]. Epoch number effects both complexity and training. Less epoch number means reduction in duration of the process but may result in a less recognition rate. In this study instead of fully training, the network is partially trained. In this Chapter it is proposed that without fully training the ANN, recognition can be done [61]. Based on the empirical observations during thesis, in a partially trained ANN, the outputs of the hidden layer may be used as a feature due to the fact that *hidden layers of an ANN learns to record the input data* so that their outputs indeed represent the input in a reduced size. Without fully training the ANN, hidden layer outputs together with Euclidean distance establishes meaningful method to identify the facial expressions. In this proposed approach, Euclidean distance is used as a classifier. Therefore the method is simple and fast compared to other well-known classifiers.

4.2. Proposed Methodology

This architecture is composed of three layers MLP. ANN is used as a tool to get the input weight data of hidden layer. It is proposed that the hidden layer output data is meaningful enough to cluster each expression. The aim herein is to partially train the ANN via dataset to get the weight matrix between input layer and hidden layer. With another words, ANN is used as a tool to extract the features of the inputted data. By using logistic sigmoid function as activation function in hidden layer ANN MLP is partially trained. By partially training, ANN cannot be able to classify the training data but by this process it is possible to get the input hidden weight vectors, then hidden layer output data can be easily calculated. It is already mentioned in Part 2.7 that weight update is a part of BP. In Figure 20 there is a deep look into a part of step by step BP algorithm related with hidden layer input and output weights.

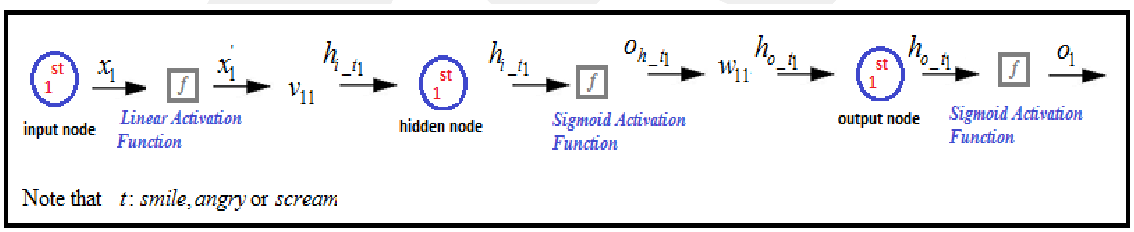


Figure 20: Calculations for first nodes' first inputs and outputs.

Consider the calculations for first nodes of MLP architecture from Figure 20, for linear activation function $x_1 = x_1^{'}$, v_{11} is the weight of the arc between 1^{st} input neuron and 1^{st} hidden layer neuron. Input enters through the hidden node is

$$h_{i\_t_1} = v_{11} x_1^{'} \tag{4-1}$$

For a logistic sigmoid transfer function, output of the 1st hidden neuron is given by

$$o_{h\_t_1} = \frac{1}{1+e^{-\lambda(h_{i\_t_1} - \Theta_h)}} \tag{4-2}$$

where Θ_h is the threshold value of 1st node of the hidden layer. Output of the hidden neuron that enters through the output neuron is

$$h_{o\_t_1} = w_{11} o_{h\_t_1} \tag{4-3}$$

where w_{11} is the weight of the arc between output of the 1st hidden neuron and 1st output neuron. To calculate the output of the 1st output neuron through a logistic sigmoid function,

$$o_1 = \frac{1}{1+e^{-\lambda(h_{o\_t_1} - \Theta_o)}} \tag{4-4}$$

where Θ_o is the threshold value of 1st node of the output layer. The aim in this study is to get $h_{o\_t_j}$ values for each expression as $h_{o\_t} = [W]_{mxn} o_{h\_t}(:, j)$. In Figure 21, h_{o_smile} is given.

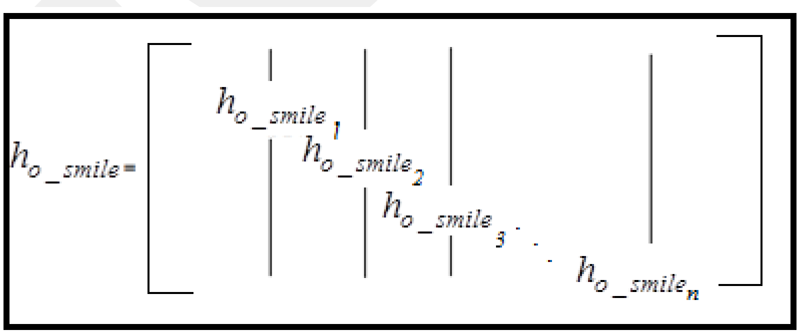


Figure 21: Hidden layer output matrix for smile class.

For an unknown test data, the same calculations of above are done to obtain the hidden layer output matrix of the test data as $h_{o\_test} = [W]_{mxn} o_{h\_test}(:, j)$.

4.2.1. Euclidean Distance

After obtaining ho_t_j matrix for each class, the centroid of these clusters $\bar{h}o\_t$ is calculated by finding mean values of each class

$$\bar{h}o\_t = \frac{1}{N} \sum_{i=1}^{N} ho\_t_i \quad t{:}smile, angry\ or\ scream \tag{4-5}$$

To decide the class of an unknown image, the Euclidean distance D between ho_test and $\bar{h}o\_t$ is calculated for each cluster (3-6). ho_test has to belong to the class in which the D value is minimum.

$$D = \left\| \bar{h}o\_t(:, i) - ho\_test(:, j) \right\|^2 \tag{4-6}$$

As it is seen in Figure 22 the ho_test belongs to first class -smile- because D is the minimum between ho_test and ho_smile.

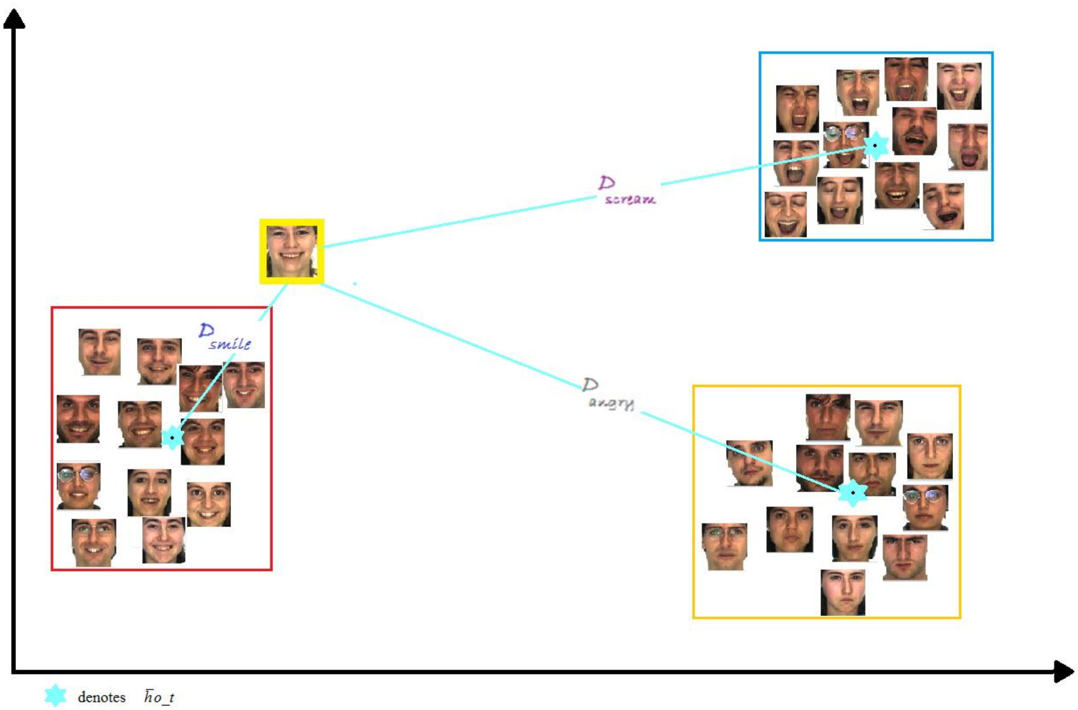


Figure 22: Proposed clustering.

4.3. Scenario 1

AR database has a handicap. In Figure 23, sample neural and angry images are given. It is obvious that even by human eyes they cannot be easily distinguished from each other.

Figure 23: Samples from AR database.

To achieve more accurate results neutral images may be omitted from the dataset. Similar problem is observed in JAFFE database due to confusion between smile and neutral expressions [14]. It is clear that truly separation of neutral and angry expressions is not possible even by human perception. Omitting neutral expressions may cause an increase in the success rate because there won't be any confusion with angry expression. Neutral image of a person can be thought as a basis since expressions are formed from neutral mood. For emotion identification, neutral image may not be a must to recognize. To validate the idea of omitting neutral images, the proposed method is applied on four expressions including neutral. 25 different male with four expressions of each are used with different hidden layer node numbers and epoch values. To enhance the images in order to increase the global contrast some basic image processing methods are used such as black white conversation, gray scale conversation and histogram equalization. In Figure 24 and 25, sample images from black white image (BWI), binary image with histogram equalization (HEBWI), gray scale image (GSI) and gray scale image with histogram equalization (HEGSI) are given.

(a) (b)

Figure 24: BWI (a) and HEBWI (b).

(a) (b)

Figure 25: GSI (a) and HEGSI (b).

Hidden nodes number related with the mapping ability of the ANN. As the number of hidden nodes increase, the network becomes more powerful. However, if this number is too large then this generalization fails. Hidden node number should fit to the size of the training data. Epoch number should be large enough to validate the minimum error. Hidden layer node number, epoch number and preprocessing methods are the factors that affect the success. But in the literature there is not an exact formulation for these parameters.

Due to empirical observations during this thesis, different combinations of these parameters (preprocess method, hidden node numbers and epoch numbers) are applied and the best results are recorded in Table 7.

Table 7: Performance results for scenario1

Preprocess method	Number of inputs	Number of hidden nodes	Number of epochs	Success rate %
BWI	100	100	200	29
HEBWI	100	100	200	27
GSI	100	75	150	27
GSI	100	75	75	31
GSI	100	90	150	26
HEGSI	100	75	100	21

Confusion effect is obviously seen in the results of Table 7 because highest recognition rate is calculated as 31% which means a fail. Aim of this scenario is to support the idea of using three facial expressions (smile, angry and scream) of AR dataset and according to the results it is clear that this idea is validated.

4.4. Scenario 2

In this scenario both in the test and training sets, 25 distinct male images are used with three expressions for each (smile, angry and scream) and the same procedure with Scenario 1 is applied. In binary conversation for each pixel there are only two possible values, 0 or 1 with no gray information. Some features that are kept in gray scale range would be smoothed to either 0 or 1 so that loss of features affect the success of the test set results. Although HE is a contrast enhancement technique and used to obtain a new enhanced image with a uniform histogram, only gray scale converted images have a better success rate in test set. As you can see from Table 8 the best combination is achieved when there are 50 nodes in hidden unit with 200 epochs and only gray scale conversion is applied. Although the ANN does not fully trained, 92% success achieved in test set by using hidden layer output nodes as feature extractor and Euclidean distance as a classifier.

Table 8: Performance results for scenario 2

Preprocess method	Number of inputs	Number of hidden nodes	Number of epochs	Success rate %
BWI	75	100	100	55
BWI	75	50	100	61
BWI	75	75	150	73
GSI	75	100	700	77
GSI	75	150	700	68
GSI	75	100	100	67
GSI	75	50	100	60
GSI	75	75	150	85
GSI	75	50	200	92
HEBWI	75	50	100	40
HEBWI	75	75	150	75
HEBWI	75	50	200	79
HEGSI	75	50	200	91
HEGSI	75	100	100	77
HEGSI	75	50	100	77
HEGSI	75	75	150	81

4.5. Scenario 3

After getting satisfactory results in Scenario 2 the input numbers for both test set and the training set are increased and woman images are included. There are 150 images in both the train set and the testing set. Regarding the success rate in GSI only conversion to gray scale is applied in this scenario. The maximum number of epoch is set to 150. In 150 epochs the ANN does not fully trained but as it is clear from the results for 100 hidden nodes and 150 epochs the success of the test set is 93% as it is seen in Table 9.

Table 9: Performance results for scenario 3

Preprocess method	Number of inputs	Number of hidden nodes	Number of epochs	Success rate %
GSI	150	75	45	63
GSI	150	75	60	60
GSI	150	75	100	77
GSI	150	100	100	69
GSI	150	100	150	93
GSI	150	100	75	59
GSI	150	150	100	79

4.6. Conclusion

In this study emotion classification by using hidden layer output data is proposed. Compared with the usual ANN applications the duration is reduced by the method of partially training the ANN to obtain hidden layer outputs. The reason why using ANN is to extract hidden layer outputs which are used as feature data. Together with Euclidean distance, successive classification with a rate of 93% is achieved.

CHAPTER 5

FACIAL FEATURES BASED DESIGN

5.1. Introduction

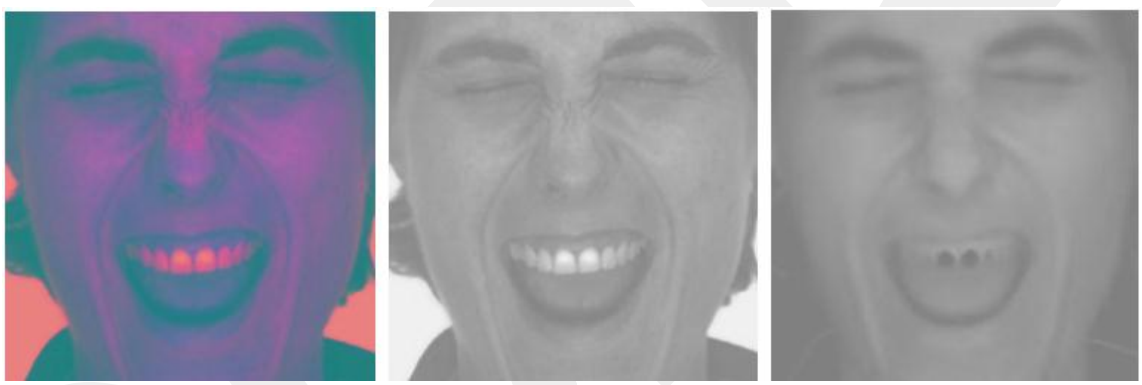
In this study, firstly skin of a facial image is detected. This detection provides a clearer appearance for the facial features such as eye, brow and mouth. By manually cropping these features as left eye, right eye and mouth, the input data of an ANN is obtained. By using BP algorithm the recognition of facial expressions is processed.

5.2. Skin Detection

Skin detection is a process to find skin colored pixels from an image. Due to the illumination conditions the skin appearance may change. The human skin color has a restricted range of hues and is not deeply saturated, since the appearance of skin is formed by a combination pigments as blood (red) and melanin (brown, yellow). For skin color modeling different approaches has been introduced [62], [63] and [64]. In this study the images from the dataset are in RGB format originally. But mixing of chrominance and luminance data make RGB not a preferable scale for color analysis and color based recognition algorithms [64]. These RGB images are converted into HSV format to extract hue and saturation values. An explicit discrimination between chrominance and luminance can be obtained by using HSV transform. For this reason HSV is a preferable color space in skin detection. In Figure 26 RGB, hue and saturated images are presented.

(a) (b) (c)

Figure 26: RGB (a), hue (b) and saturated (c).



(a) (b) (c)

Figure 27: YUV (a), C_b (b) and C_r (c).

Then RGB image is converted into YC_rC_b model. To identify the skin portion of an image, C_b and C_r values of RGB image are calculated. In Figure 27, YUV, C_b and C_r converted images are shown. After evaluating hue, saturation, C_r and C_b values, in a range of threshold, these values are checked even if they satisfy the threshold conditions or not. If they do, then it is said to be a skin region; else non skin. As it is stated in Equation (2-31) the skin color pixels should satisfy the conditions of 0 <= ***hue*** <= 0.25; 0.15 <= ***saturation*** <= 0.9 and in Equations (2-32) and (2-33) the skin portion of an image should satisfy the conditions of 140 <= C_r <= 165; 140 <= C_b <= 195. For all train and test images the above 4 conditions are held. As a result Figure 28 is generated. In this figure skin region are represented by 1 and non-skin regions are represented by 0.

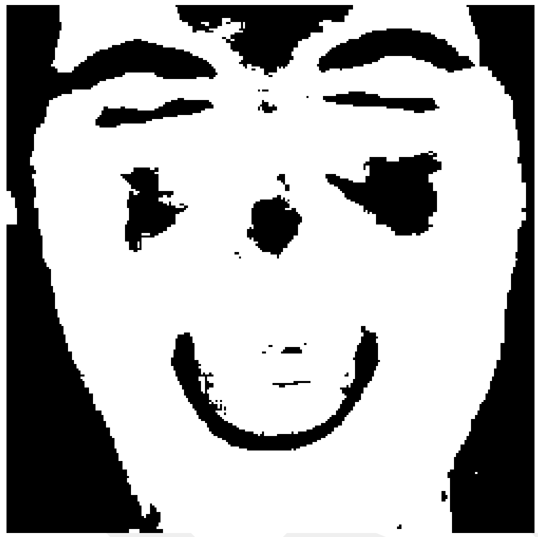


Figure 28: Skin detected image.

By applying skin detection the features which are needed for facial expressions appear more visible and by this way redundant data of a face is eliminated so that eyes, brows and mouth of each image are extracted.

5.3. Cropping

The aim herein is to crop each eye+brow and mouth from an image. These values are held in a vector then inputted through the ANN. The cropped features are resized with a ratio of 1/10 ratio. The cropping scales of each feature are defined manually due to empirical observations on AR images. Since the largest change interval for a feature occurs in scream expressions, this expression makes the basis of cropping scale. By using the *imcrop* function of Matlab, which is given in (5-1), cropping is done. This function takes two arguments: image to crop and rectangle coordinates. This rectangle has four elements position vector, *[xmin ymin width height]*, that specifies the size and position of the crop rectangle. In Figure 29, sample cropped features of an image are given.

```
Lefteye = imresize(imcrop(img2, [15 30 105 85]),1/10);
Righteye = imresize(imcrop(img2, [120 30 105 85]),1/10);        (5-1)
Mouth = imresize(imcrop(img2, [40 170 175 59]),1/10);
```

where *img2* is the image which is to cropped, is followed by four element position vector that holds *xmin,ymin*, *width* and *height* that specifies the position and size of the crop rectangle and 1/10 rescalation is applied on the cropped images.

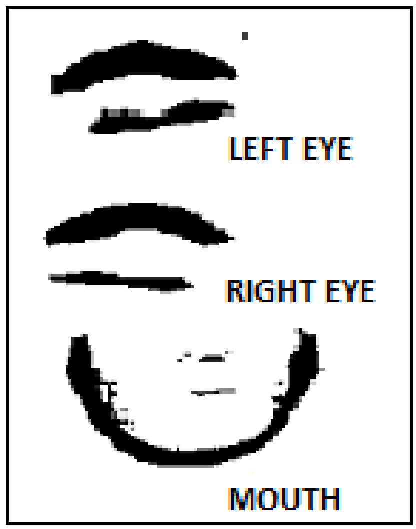


Figure 29: Cropped features.

These three vectors are concatenated and this process is applied for each image of train set and the test set.

5.4. Architecture

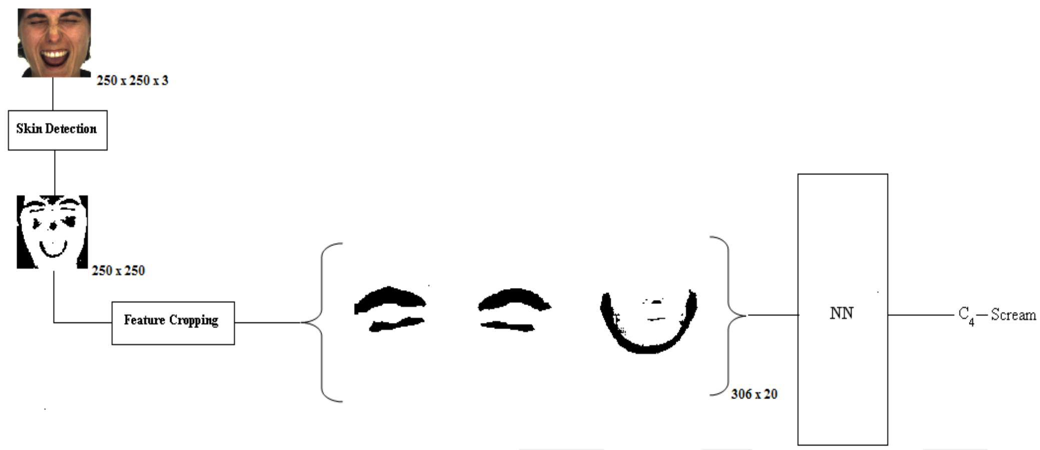


Figure 30: Proposed architecture.

In Figure 30, the whole system architecture is presented. In the first step, the skin detection is applied to the RGB images from AR dataset. After detecting skin, the rest of the features (brow, eye, noise, cheek and mouth) seemed clearer. For facial expressions identification brow+eyes and mouth are the necessary features. Only dealing with three of them for each image is meaningful in case of studying with precision data only. To extract these features from skin and other redundant features of a face, image cropping is applied. Average cropping parameters are calculated empirically and the features are cropped as left eye, right eye and mouth. These cropped images are putted in a vector which forms the input layer of ANN. Inside ANN BP algorithm is used with logistic sigmoid function of activation function. As an output of ANN, classification for each expression is obtained.

5.5. Results

For 40 men and 35 women with four expressions of each there are totally 300 inputs were trained by the proposed architecture and 100 % success is achieved in training set. For test set 11 men and 11 women images (88 test inputs) are used. For different number of hidden nodes with a fix epoch number of 500, the architecture is trained. The results are shown in Table 10 with the most successful rate of 67%.

Table 10: Performance results for four expressions for 500 epochs

Number of epochs	Number of hidden nodes	Success rate %
500	50	65
500	75	64
500	100	67
500	120	66
500	150	66
500	170	67

By considering the fact of confusion of neutral image with angry as it was stated in Chapter 3 and 4, neutral images are omitted and the same procedure is applied for 225 train and 66 test data. As it is seen in Table 11 the success rate is increased by 13% and 80% overall success rate is obtained.

Table 11: Performance results for three expressions for 500 epochs

Number of epochs	Number of hidden nodes	Success rate %
500	50	76
500	75	80
500	100	74
500	120	77
500	150	79
500	170	67

For 750 epochs, same number of data and same combinations of hidden nodes are applied. For 270 epochs, 170 hidden nodes 83% success is performed as it is given in Table 12.

Table 12: Performance results for three expressions for 750 epochs

Number of epochs	Number of hidden nodes	Success rate %
750	50	80
750	75	77
750	100	75
750	120	77
750	150	79
750	170	83

5.6. Conclusion

In this study, skin detection is applied to RGB images as a preprocessing method. By detecting skin, facial features are also extracted. To extract brow+eyes and mouth from other features, manual cropping is done. Then these cropped features are used as input of ANN. The handicap of dataset is observed also in this study while studying with four expressions. Therefore neutral expressions are eliminated. By using facial features based method with ANN, facial expressions identification is obtained with 83% recognition rate.

CHAPTER 6

CONCLUSION AND FUTURE WORKS

6.1. Conclusion

The objective of this thesis project is to investigate different approaches for ANN based facial expressions identification. For this purpose three successful approaches are proposed. These approaches give the best recognition rates when neutral images are omitted.

First, tree based ANN is designed to increase the performance of architecture due to the confusion between angry and neutral expressions. In this method three PCA+ANN architectures are used in a cascaded manner where PCA is used to reduce the dimensions of the input images of ANN. By applying this architecture, 88.30 % recognition rate is obtained.

Second, emotion identification using hidden layer outputs is proposed. The aim of this approach is to partially train the ANN via dataset to get the weight matrix between input layer and hidden layer. ANN is used as an extractor for the features of the input data. Without fully training, the duration of the process is reduced and together with Euclidean distance as a classifier, a successful classification rate of 93% is achieved.

Finally, facial features based design of ANN is proposed. For this study skin detection is applied to RGB images. After detecting the skin, facial features seemed more apparently. For emotion identification brow, eye and mouth features of an image are enough to use. To extract these features, brow+eyes and mouth of each image are cropped and then inputted through ANN. As a result of this method 83 % recognition is achieved.

6.2. Future Works

In this thesis AR dataset is used for identification of facial expressions. By using different facial expression datasets, such as JAFFE, the proposed methodologies may give better results.

The most successful method is given in Chapter 4, for 150 images by applying GSI conversation, with 100 hidden nodes and 150 epochs 93% success is obtained. By applying different combinations of input numbers, epoch numbers and hidden node numbers around the best combination a better result may be achieved.

In Chapter 5, manual cropping is applied to the images. By providing an automatic cropping mechanism more accurate cropping may be held so the success of ANN may increase due to the new input data.

REFERENCES

[1] H. Halawani, “Emotion recognition system based on facial expression,” M.S. Thesis, University of Sheffield, 2011.

[2] B. Azar, “What’s in a face?,” American Psychology Association, vol. 31, No. 1, 2000, URL http://www.apa.org/monitor/jan00/sc1.aspx

[3] B. Fasel, and J. Luettin, “Automatical facial expressions analysis: a survey,” Pattern recognition 36, pp. 259–275, 2003.

[4] P. Ekman, and W. V. Friesen, “Constants across cultures in the face and emotion,” J. Personality Social Psychology 17 (2), pp. 124–129, 1971.

[5] M. Suwa, N. Sugie, and K. Fujimora, “A preliminary note on pattern recognition of human emotional expression,” Proc. of the 4^{th} Int. Joint Conf. on Pattern Recognition, Kyoto, Japan, pp. 408–410, 1978.

[6] K. Mase, and A. Pentland, “Recognition of facial expression from optical flow,” IEICE Trans. E 74 (10), pp. 3474–3483, 1991.

[7] F. Bourel, C. C. Chibelushi, and A. A. Low, “Robust facial expression recognition using a state-based model of spatially-localised facial dynamics,” Proc. 5^{th} IEEE Int. Conf. Automatic Face and Gesture Recognition, pp. 106–111, 2002.

[8] P. Ekman, and W. V. Friesen, “Facial action coding system: a technique for the measurement of facial movement,” Consulting Psychologists Press, Palo Alto, 1978.

[9] H. Kobayashi, and F. Hara, “Recognition of six basic facial expressions and their strength by neural network,” Proc. IEEE Int. Workshop on Robot and Human Communication, New York, pp. 381–386, 1992.

[10] L. Ma, and K. Khorasani, “Facial expression recognition using constructive feed forward neural networks,” IEEE Trans. on Systems, Man and Cybernetics, Part B: Cybernetics, vol. 34, No. 3, 2007.

[11] J. Y. Chang, and J. L. Chen, “Automated facial expressions recognition system using neural networks,” Journal of Chinese Inst. of Engineers, vol. 24, No. 3, pp. 345–356, 2001.

[12] G. U. Kharat, and S. V. Dudul, “Emotion recognition from facial expressions using neural networks,” Advances in Soft Computing, vol. 60, pp. 207–219, 2009.

[13] D. T. Lin, “Facial expression classification using PCA and Hierarchical Radial Basis function network,” Journal of Information Science and Engineering, vol.22, pp. 1033–1046, 2006.

[14] P. Lekshmi, and M. Sasikumar, “A neural network based facial expression analysis using Gabor wavelets,” World Academy of Science, Engineering Technology 42, 2008.

[15] J. J. Lien, T. Kanade, J. F. Cohn, and C. C. Li, “Automated facial expression recognition based on FACS action units,” Proc. 3th IEEE Int. Conf. Automatic Face and Gesture Recognition, pp. 390–395, 1998.

[16] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, No. 10, pp. 974–989, 1999.

[17] A. Sayeed, M. Sohail, and P. Bhattacharya “Classification of facial expressions using k-nearest neighbor classifier,” A. Gagalowicz and W. Philips (Eds), MIRAGE, LNCS 4418, pp. 555–556, 2007.

[18] T. Otsuka, J. Ohya, and R. Nakatsu, “Recognizing facial expressions generated by multiple persons based on Hidden Markov Models with continuous output probability densities,” The Trans. of the Institute of Electronics, Information and Communication Engineer D-II, vol. J80-D-II, No. 8, pp. 2129–2137, 1997.

[19] K. Mase, “Recognition of facial expression from optical flow,” Institute of Electronics, Information and Communication Engineers Transactions, vol. E74, pp. 3474–3483, 1991.

[20] M. Rosenblum, Y. Yacoob, and L. S. Davis, “Human emotion recognition from motion using a Radial Basis Function network architecture,” University of Maryland, Technical Report CS-TR-3304, 1994.

[21] Y. Yacoob, and L. Davis, “Recognizing human facial expression,” University of Maryland, Technical Report CS-TR-3265, 1994.

[22] H. Kobayashi, and F. Hara, "The recognition of basic facial expressions by neural network," Proceedings of Int. Joint Conf. on Neural Network, pp. 460–466, 1991.

[23] H. Kobayashi, and F. Hara, “Recognition of six basic facial expressions and their strength by neural network,” IEEE Int. Workshop on Robot and Human Communication, pp. 381–386, 1992.

[24] H. Kobayashi, and F. Hara, "Recognition of mixed facial expressions by neural network," IEEE Int. Workshop on Robot and Human Communication, pp. 387–391, 1992.

[25] J. Ding, M. Shimamura, H. Kobayashi, and T. Nakamura, "Neural network structures for expression recognition," Proceedings of Int. Joint Conference on Neural Network, pp. 1420–1423, 1993.

[26] T. Hashiyama, and T. Furuhashi, Y. Uchikawa, and H. Kato, "A face graph method using a fuzzy neural network for expressing conditions of complex systems," Neural Networks IEEE Int. Joint Conference on Digital Object Identifier, vol. 2, pp. 1600–1605, 1991.

[27] A. Ralescu, and R. Hartani, "Some issues in fuzzy and linguistic modeling," IEEE Proc. of Int. Conf. on Fuzzy Systems, 1995.

[28] H. Ushida, T. Takagi, and T. Yamaguchi, "Recognition of facial expressions using conceptual fuzzy sets," Proc. of the 2nd IEEE Int. Conf. on Fuzzy Systems, pp. 594–599, 1993.

[29] T. Sato, and T. Yamaguchi, "Generation of facial expression using chaotic retrieval," IEEE Symposium on Emerging Technologies and Factory Automation, pp. 213–128, 1994.

[30] G. Cottrell, and J. Metcalfe, "EMPATH: face, gender, and emotion recognition using hoons," in J. Moddy, R. P. Lippman, and D. S. Touretzky, (eds.), Advances in Neural Information Processing Systems, vol. 3, pp. 564–571, 1991.

[31] C. Padgett, and G. Cottrel, "Identifying emotion in static face images," in The 2nd Joint Symposium of Neural Computation, vol. 5, pp. 91–101, 1995.

[32] K. Matsuno, C. Lee, and S. Tsuji, "Recognition of human facial expressions without feature extraction," in Proc. of the European Conference on Computer Vision, pp. 513–520.

[33] Gharate M.K, and R. A. Patil, "Neural network," Pharmaceutical Reviews, vol. 5, issue 6, 2007.

[34] M. Negnevitsky, "Artificial intelligence: a guide to intelligent systems," Addison Wesley, 2001.

[35] S. Tartarotti, "Neuron Cell", URL http://www.creativecrash.com/marketplace/3d-models/anatomy/internal/c/neuron-cell, 2006.

[36] J. Kutsurelis, "Forecasting financial markets using neural networks: an analysis of methods and accuracy," Master Thesis, Naval Postgraduate School, 1998.

[37] T. Jackson, and R. Beale, “Introduction to neural networks,” Adam Hilger (eds), Bristol, Philadelphia and New York, 1990.

[38] M. T. Hagan, H. B. Demuthand, and M. H. Beale, “Neural network design,” Boston, MA: PWS Publishing, 1996.

[39] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain,” Psychological Review, pp. 386–408, 1958.

[40] M. Minskey, and S. Papert, “Perceptrons,” MIT Press, Cambridge, 1969.

[41] R. Duda, P. Hart, and D. Stork, “Pattern classification”, 2nd Edition, Wiley Interscience, New York, 2001.

[42] J. Nazzal, I. El-Emary, and S. Najim, “Multilayer perceptron neural network (MLPs) for analyzing the properties of Jordan oil shale,” World Applied Sciences Journal5, pp. 546–552, 2008.

[43] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feed forward networks are universal approximators,” Neural Networks, vol. 2, pp. 359–366, 1989.

[44] G. Cybenko, “Approximation by superposition of a sigmoidal function,” Mathematics of control, Signal and Systems, pp.303–314, 1989.

[45] P. Phillips, H. Moon, P. Rauss, and S. Rizvi, “The FERET evaluation methodology for face recognition algorithm”, IEEE Trans. PAMI, 2000.

[46] T. Sim, S. Baker, and M. Bsat, “The CMU pose, illumination and expression database,” IEEE Trans. PAMI, 2003.

[47] M. Lyons, J. Budynek, and S. Akamatsu, “Automatic classification of single facial images,” IEEE Trans. PAMI, pp. 1357–1362, 1999.

[48] H. Deng, L. Jin, L. Zhen, and J. Huang, “A new facial expression recognition method based on local Gabor filter bank and PCA plus LDA,” Int. Journal of Information Technology, vol.11, No. 11, 2005.

[49] A. Martinez, “The AR face database,” URL http://www2.ece.ohiostate.edu/~aleix /ARdatabase.html.

[50] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, “A high-resolution 3D dynamic facial expression database,” IEEE 8th Int. Conf. on Automatic Face and Gesture Recognition, pp. 1–6 , 2008

[51] M. Grgic, and K. Delac, “Face recognition homepage,” http://www.face-rec.org/

[52] M. Turk, and A. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, vol. 3, No. 1, 1991.

[53] I. K. Fodor, "A survey of dimension reduction techniques," LLNL technical report, 2002.

[54] R. C. Gonzales, and R. E. Woods, "Digital image processing," Prentice-Hall, Inc., Englewood Cliffs, 2008.

[55] A. R. Robertson, "Color perception," Physics Today, vol. 45, pp. 24–29.

[56] S. Chitra, and G. Balakrishnan, "Comparative study for two color spaces HSCbCr and YCbCr in skin color detection," Applied Mathematical Sciences, vol. 6, No. 85, pp. 4229–4238, 2012.

[57] H. Tora, and M. A. Günler, "Tree based neural network design for emotion identification analysis," IEEE 19th Conf. on Signal Processing and Communications Applications, SIU, pp. 881–883, 2011.

[58] N. Rose, "Facial expression classification using Gabor and Log-Gabor filters," IEEE an Automatic Face and Gesture Recognition, pp. 346–350, 2006.

[59] Y. Saatci, and C. Town, "Cascaded classification of gender and facial expression using active appearance models," IEEE 7th Int. Conf. on Automatic Face and Gesture Recognition, pp 393–400, 2006.

[60] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Cernocky, "Strategies for training large scale neural network language models," Automatic Speech Recognition and Understanding (ASRU), pp. 196–201, 2011.

[61] M. A. Günler, and H. Tora, "Emotion classification using hidden layer outputs," International Symposium on Innovations in Intelligence Systems and Applications (INISTA), pp. 1–4, 2012.

[62] B. D. Zarit, B. J. Super, and F. K. H. Quek, "Comparison of five color models in skin pixel classification," ICCV'99 Int'l Workshop on recognition, analysis and tracking of faces and gestures in Real-Time systems, pp. 58–63, 1999.

[63] J. C. Trerillon, M. N. Shirazi, H. Fukamachi, and S. Akamatsu, "Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images," Proc. of the Int. Conf. on Face and Gesture Recognition, pp. 54–61, 2000.

[64] J. Brand, and J. Mason, "A comparative assessment of three approaches to pixel level human skin detection," Proc. of the Int. Conf. on Pattern Recognition, vol. 1, pp. 1056–1059, 2000.