**THE REPUBLIC OF TURKEY**
**BAHÇEŞEHİR UNIVERSITY**

# FACE DETECTION AND FACIAL EXPRESSION RECOGNITION USING MOMENT INVARIANTS

**Master Thesis**

**ALİ KARAALİ**

**ISTANBUL, 2012**

**THE REPUBLIC OF TURKEY**
**BAHÇEŞEHİR UNIVERSITY**


**THE GRADUATE SCHOOL OF NATURAL AND APPLIED**
**SCIENCES ELECTRICAL AND ELECTRONICS ENGINEERING**


# FACE DETECTION AND FACIAL EXPRESSION
# RECOGNITION
# USING MOMENT INVARIANTS


**Master Thesis**


**Ali KARAALİ**


**Supervisor: Assoc. Prof. Çiğdem EROĞLU ERDEM**


**ISTANBUL, 2012**

**THE REPUBLIC OF TURKEY**
**BAHÇEŞEHİR UNIVERSITY**


**THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES**
**ELECTRICAL AND ELECTRONICS ENGINEERING**

Title of the Master's Thesis: Face Detection and Facial Expression Recognition Using Moment
Invariants
Name/Last Name of the Student: Ali KARAALİ
Date of Thesis Defense: 26.01.2012


The thesis has been approved by the Graduate School of Natural and Applied Sciences.


Assoc. Prof. Tunç BOZBURA
Director


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


Assoc. Prof. Ufuk TÜRELİ
Program Coordinator


This is to certify that we have read this thesis and that we find it fully adequate in scope, quality
and content, as a thesis for the degree of Master of Science.


| Examining Committee Members | Signatures |
|---|---|
| _____ | _____ |
| Assoc. Prof. Çiğdem EROĞLU ERDEM (supervisor) | _____ |
| Asst. Prof. Devrim ÜNAY | _____ |
| Asst. Prof. Olcay KURŞUN | _____ |

# ACKNOWLEDGEMENTS

First of all I would like to thank my thesis supervisor Assoc. Prof. Çiğdem EROĞLU ERDEM, who has given me the opportunity to work on this thesis. I'm very grateful for her support, guidance, encouragements and invaluable help during the preparation of this thesis.

I would like to thank Dr. H. Fatih UĞURDAĞ who encouraged me during my master studies, and on my Master Thesis. I would also like to thank Dr. Olcay KURŞUN for his contribution of the last chapter of this thesis.

My special thanks to my friends and colleagues, especially to team members for their comments, suggestions and their endless support all through this work, also in my personal life and masters degree courses.

Last but not the least; I would like to express my love and gratitude to my family for their support, motivation and patience.

January 26, 2012                                                                                    Ali KARAALİ

# ABSTRACT

FACE DETECTION AND FACIAL EXPRESSION RECOGNITION USING MOMENT
INVARIANTS

Ali Karaali

Electical and Electronics Engineering
Supervisor: Assoc. Prof. Çiğdem EROĞLU ERDEM

January 2012, 92 pages

The problem of face detection refers to determining whether or not there are any faces in a given image and estimating the location and size of any detected faces. Face detection is an important first step of many computer vision applications including but not limited to security, face recognition, photography, and facial expression recognition. Face detection is a trivial task for humans, however it is not very easy for computers due to geometric (scale, in-plane rotation, pose, occlusions etc.) variations. In this thesis, we propose a new efficient technique for localization of faces in arbitrary images. The technique is based on segmentation of images into skin colored blobs, which is followed by computation of scale, translation and rotation invariant moment-based features to learn a statistical model of faces and non-faces. Our method is trained and evaluated on a standard, publicly available face image database and its performance is assessed over a range of statistical pattern classifiers. The superiority of the method to the state-of-the art face detection methods is its ability to detect non-frontal faces in a person-independent way. Experimental results show that the proposed method gives higher true positive rates as compared to the well known Viola-Jones face detector.

One of the other important problems of computer vision and image processing is recognition of human emotions from facial expressions. Although emotions are recognized by humans easily, it is not that easy using computer vision and image processing algorithms. In this thesis, we proposed two facial expression recognition algorithms. The first algorithm is a rotation invariant facial expression recognition method using Zernike moment invariants. We used a sparse representation-based classifier to discriminate the facial expressions. The system is fully automatic, person independent and can recognize six different facial expressions. In the second algorithm; we show that the sparse representation of expressive facial images from a basis set of neutral face images can be utilized in facial expression recognition. The method consists of face

and facial expression recognition parts and uses a two stages sparse representation-based classifier. In the first part of this method, residual images are obtained during the face recognition step  which contains the facial expression related components of the expressive face mostly, and then the expression in the residual images are classified using a difference image dictionary. Experimental results are conducted on Extended Cohn-Kanade (CK+) database and results of the proposed algorithms are promising.

**Keywords:** Moment Invariants, Sparse Representation-Based Classification, Face Detection, Facial Expression Recognition, Zernike Moments

# ÖZET

## MOMENT ÖZNİTELİKLERİNİ KULLANARAK YÜZ KONUM TESPİTİ VE YÜZSEL İFADE TANIMLAMA

Ali Karaali

Elektrik Elektronik Mühendisliği Yüksek Lisans Programı
Tez Danışmanı: Doç. Dr. Çiğdem Eroğlu Erdem

Ocak 2012, 92 sayfa

Yüz konum tespiti problemi verilen bir imge üzerinde yüz olup olmadığının anlaşılması ve var olan yüz bölgelerinin doğru konum ve büyüklükte tespit edilmesi olarak tanımlanmaktadır. Yüz konum tespiti güvenlik, yüz tanımlama, yüzsel duygu kestirimi gibi bir çok bilgisayarlı görü problemlerinde bir ön evre olarak kullanılmaktadır. İnsan açısından yüz konum tespiti çok basit bir işlem olarak kabul edilsede, boyut değişimi, yatay düzlemde açı değişimi ve yüzü kapatan cisimlerin olaması gibi durumlardan dolayı bilgisayarlı görü teknikleri ile yüz konum tespiti işlemi çözülmesi kolay bir problem değildir. Bu tez çalışmasında, imge üzerindeki yüz bölgelerinin konumunun bulunması için yeni bir algoritma önerilmiştir. Önerilen bu algoritma ten rengi bölütlemesiyle elde edilen imge üzerinde ki ten rengi bölgelerinden boyut değişimi, öteleme ve dönme gibi geometrik dönüşümlerden etkilenmeyen öznitelik vektörlerinin istatistiksel bir modele dayalı olarak eğitilmesine dayanmaktadır. Önerilen algoritma standart ve oldukça yaygın kullanılan yüz imgelerinden oluşan veritabanında denenip başarımı analiz edilmiştir. Önerilen methodun diğer yaygın olarak kullanılan methodlardan üstünlüğü, imge üzerindeki yüz bölgelerini ön cepheden olmayan durumlarda ve kişiden bağımsız olarak tespit edebilmesidir. Yapılan deneysel sonuçlar göstermiştir ki Viola-Jones gibi yaygın kullanılan yüz konum tespiti algoritmasını doğru bulma açısından geçmektedir.

Bilgisayarlı görü ve imge işleme alanlarının sahip olduğu bir diğer problem de insan duygularının yüz ifadelerinden algılanmasıdır. İnsanların ifade ettiği temel yüzsel duygular insanlar tarafından kolayca algılanabilesine rağmen, bu tanımlama işlemi algoritmik açıdan kolay değildir. Bu tez çalışmasında iki adet yüzsel ifade tanımlama algoritması önerilmiştir. Önerilen duygu tanımlama algoritmalarının ilki tespit edilecek yüz ifadesinin koordinat düzlemiyle yaptığı açıdan etkilenmeyen Zernike Momentleri ile sağlanmıştır. Elde edilen Zernike momentleri daha sonra hangi duygunun içerdiği seyreltik betimleme tabanlı bir sınıflandırıdan (SBS) geçirilerek sağlanmıştır. Önerilen sistem tamamen otomatik olup, kişiden bağımsız olarak 6 adet yüzsel ifadeyi tanımlayabilmektedir. Önerilen diğer methot ifadesel yüz imgelerinin nötr imge tabanlı bir kümeden seyreltik betimlenmesi ile yüzsel ifadelerin tanımlanabilinmesi üzerinedir. Methot, iki aşamalı seyreltik betimleme tabanlı sınıfladırıcı

kullanımıyla yüz ve yüzsel ifade tanımlanmasından oluşmaktadır. Önerilen bu methodun ilk safhasında, yüz tanımlama probleminin SBS ile çözümü sayesinde elde edilen ve çoğunlukla yüzdel ifade içeren rezidüel imgeler, fark imgelerinden oluşan bir eğitim kümesi içinde seyreltik betimleme tabanlı sınıflandırılmıştır. Genişletilmiş Cohn-Kanade (CK+) veri tabanı üzerinde yapılan deneysel sonuçlar göstermektedir ki önerilen method oldukça gürbüzdür.

**Anahtar Kelimeler:** Moment Değişmezleri, Seyreltik Betimleme Tabanlı Sınıflandırıcı, Yüz Konum Tespiti, Yüzsel İfade Kestirimi, Zernike Momentleri

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

AAM    : Active Appearance Model

CK+    : Extended Cohn Kanade Database

CRT    : Cathode Ray Tube

CVL    : Computer Vision Laboratory

DCT    : Discrete Cosine Transform

FACS    : Facial Action Coding System

FDDB    : Face Detection Database

FER    : Facial Expression Recognition

FP    : False Positive

GPU    : Graphical Processing Unit

HCI    : Human Computer Interaction

KNN    : K Nearest Neighbor

LB    : Linear Bayes

LUT    : Look Up Table

LOSO    : Leave One Subject Out

MR    : Miss Rate

NN    : Nearest Neighbor

NS    : Nearest Subspace

OG    : Orthogonal

RGB    : Red Green Blue

QB    : Quadratic Bayes

SPM    : Skin Probability Map

SRC    : Sparse Representation-Based Classifier

SVM    : Support Vector Machines

TP    : True Positive

VJ    : Viola Jones

# LIST OF SYMBOLS

| | | |
|---|---|---|
| Luma, Blue Difference, Red Difference | : | $YC_rC_b$ |
| Mean | : | $\mu_s$ |
| Covariance Matrix | : | $\Sigma_s$ |
| Number of Skin Pixels | : | $T_s$ |
| Number of Non-skin Pixels | : | $T_n$ |
| Area of a Skin Color Region | : | $A_s$ |
| Image Function | : | $f(\rho, \theta)$ |
| $\ell^1$ Norm | : | $\|x\|_1$ |
| $\ell^2$ Norm | : | $\|x\|_2$ |

# 1.  INTRODUCTION

Gestures are expressive, meaningful body motions involving physical movements of face, head or body with the intent of; deploying meaningful knowledge and interacting with environment. Amongst the gestures that are made by humans, facial gestures are relatively important because they involve extracting sensitive features such as emotional states and they have many applications in our daily lives. Besides, facial expressions play an important role in the cognition of human emotion and communication. Basic facial expressions such as happiness, sadness, fear, anger, disgust and surprise (Ekman et al. 1971) are recognized by humans regardless of ethnicity and nationality. Although individuals have different appearance of these expressions, humans have the ability which is recognizing a wide range of various expressions. For example, even though we are not familiar with someone's face, we are able to recognize the person's facial expression due to the universality of affect expressions (Ekman et al. 1971). However, it is a challenging task for a computer vision system to classify the universal facial emotion across different person.

Many theories of emotion have been proposed, and there is little agreement about a definition of emotion. Some of these definitions are verified after measurement of some physiological signal become available. Emotions are, in general, short-term. According to Darwin (1890), expression of emotion is related to survival and thus, nonverbal communication is as important as the verbal interaction. James (1890) viewed emotions not as causes but as effects. According to James, "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". At about the same time Carl Lange stated a similar theory. This is referred to as emotion's theory of "James Lange".

The other interesting and important topic is how the researchers managed to acquire data for observation. Some of the researchers used posers, including professional and non-professional actors. Some other researchers used some movies or pictures to induce emotional reaction such as an autopsy, a scene from a horror movie etc. Ekman did many experiments about the obtaining data and concluded that expressions can be convincingly portrayed (Ekman 1982).

## 1.1 MOTIVATION

Considering the growing technology it is not too hard to notice the importance of interaction between humans and computers. As computers become integrated into everyday object that we use, effective natural human-computer interaction becomes critical. In many applications users need to be able to interact naturally with computers. For example, in application where a computer takes a role such as an instructor, a helper or an assistant, it may enhance the computer's functionality to be able to know users' emotional state. Another application of a facial emotion recognition system is to help human users to monitor their stress level.

One other important application of an expression recognition system is to monitor automobile's driver alertness/drowsiness levels etc. (Vural et al. 2008) This plays an important role in the social context due to the fact that reduces the dead resulted traffic accidents.

Facial expression recognition also plays an important role in security applications such as interrogation. Not only electrical signals but also they use facial emotions to evaluate interviewee's honesty (Ryan et al. 2009).

## 1.2 CONTRIBUTION AND OUTLINE OF THE THESIS

This thesis has four contributions;

1. The first contribution of the thesis is the evaluation of four skin color detection algorithms. These skin color algorithms are analyzed mainly for the face detection problem since eliminating redundant regions of images allows researchers to develop computationally efficient algorithms; on the other hand, developing such algorithms that gives a low false positive rates while keeping a high true detection rate. We used these skin color detection techniques in a face detection algorithm and evaluated their performances.

2. A new face detection algorithm is presented, which begins with segmentation of images into skin colored blobs. These blobs are used for computation of scale, translation and rotation invariant features to learn a statistical model of faces and non-faces. Finally, the computed features are classified into face and non-face classes using several well-known classifiers. The evaluation of our algorithm shows that the proposed method gives high true detection rates as compared to other well-known face detectors such as the Viola-Jones face detector.

3. A new facial expression recognition algorithm is presented based on Zernike moment invariants and a sparse representation-based classifier (SRC) that has become very popular in recent years. Zernike moment features with SRC is compared to the original pixel intensity features under the presence of rotation. Experimental results on Extended Cohn-Kanade (CK+) database show that the accuracy of Zernike moments outperforms the original pixel intensity features.

4. In the last chapter of the thesis, we show that sparse representation of expressive images from a basis set of neutral face images can be utilized in facial expression recognition. The method is based on the idea that the non-representable part of the expressive face images can be a good approximation of the difference images when true neutral face is not known for a given expressive test image. Experimental results on the CK+ database are promising, giving a recognition rate of 82%.

The outline of the thesis as follows.

In chapter two, skin color segmentation methods are evaluated in detail and experimental results are given with numerical and visual outcomes.

In chapter three, the proposed multi view face detector is presented and the details of the algorithm are explained elaborately. Results of various moment features using several classifiers with different skin color detectors are shown.

In chapter four; first, the facial expression algorithm is presented using Zernike and original pixel intensity features. The features are investigated under the presence of rotation. Second, we show that the residual images obtained from face recognition can

be utilized as a difference image if the person is not known/given and experimental results are shown and assessed.

In the last chapter, we present the overall and we mentioned possible future research directions.

# 2. A COMPARISON OF SKIN COLOR DETECTION METHODS

Skin color is a useful cue for face detection, localization and tracking since color is highly robust to geometric variation of face and allows fast processing. Pixel level skin color detection is significantly rapid and it reduces the search space prior to other high level classifications (Gang and Sung 1999). Thus, such detectors are used in the pre-processing step of high level face detection and emotion recognition system.

Color allows researchers to develop fast algorithms due to the fact that it is computationally efficient and highly robust to geometric variations of face patterns. Also, experimental results show that human skin has a characteristic color, which is recognized by humans easily. Hence, developing a system that uses skin color as a cue for face detection was an idea suggested either by task properties or common sense.

When building a system that uses skin color as a cue for facial feature extraction, the major steps are first; selection of appropriate color space which will give more discrimination between skin and non-skin color regions, second; modeling the distribution of skin color properly and finally; selecting a way of processing the skin color segmentation results. In this thesis, we discuss all of the three questions.

## 2.1 LITERATURE SURVEY

In this thesis, we mainly discuss pixel-based skin detection methods, which classify a pixel as skin or non-skin regardless of its neighborhoods. Unlike the region based methods, it tries to consider the spatial information of skin pixels during the classification stage to induce to accuracy.

Pixel-level skin color detection has a long story, but there have been many contributions towards to skin color detection during the several past decades. Zarit et al. (1999) have published a survey includes comparison of five different techniques and two individual non-parametric skin modeling methods. Terrillon et al. (2000) have compared two parametric techniques and nine color space. Brand and Mason (2000) have analyzed three different color modeling techniques. Besides, Lee and Yoo (2002) have evaluated

two popular skin modeling techniques in different color space and proposed a model of their own.

High accuracy of a skin color detection method depends on the chosen color space; hence, here it can be found a brief review of the most popular color spaces and their properties.

### 2.1.1 Color Spaces For Skin Color Modeling

Colors have different properties from each other, and many of these color spaces have been adapted to skin color segmentation problem. Here below, you can find the most popular color spaces and their properties.

### 2.1.1.1 RGB

Originally, RGB color space originated from CRT applications that it describes the color as a combination of three color rays such as red, green and blue. This color space is wide variety of usage for processing and storing of digital data. In this color space, however, channels have a little correlated to each other and it can be a drawback in some situations. Nevertheless, many researchers such as Brand and Mason (2000) and John and Rehg (1999) used this color space to differentiate the skin and non-skin color pixels.

### 2.1.1.2 Normalized RGB

Normalized RGB can be easily obtained from the normalization of original RGB values:

$$r = \frac{R}{R+G+B} , g = \frac{G}{R+G+B} , b = \frac{B}{R+G+B} \qquad \text{(2.1)}$$

The sum of the three components equals to 1 and is called normalized channels. The third component does not carry much information and could be omitted to reduce to

dimensionality. The remaining color channel components are generally called "pure colors".

### 2.1.1.3  Hue – Saturation – Intensity (Value, Lightness)

Hue – saturation based color components is a result of the necessity to specify color properties numerically. This color space describes color with intuitive values based upon the artist's idea of tint, saturation and tone. The dominant color is described by Hue and colorfulness of an area is measured by saturation (Poynton 1995). The intensity or lightness or value is completely related to the luminance. This color space holds good benefits in a skin color classification system due to intuitiveness of color components and explicit discrimination between luminance and chrominance (McKenna et al. 1998).

$$H = \arccos \frac{\frac{1}{2}((R-G)+(R-B))}{\sqrt{((R-G)^2+(R-B)(G-B))}} \qquad \text{(2.2a)}$$

$$S = 1 - 3\frac{\min (R,G,B)}{R+G+B} \qquad \text{(2.2b)}$$

$$V = \frac{1}{3}(R + G + B) \qquad \text{(2.2c)}$$

To reduce the dependency of chrominance on the illumination level, Fleck et al. (1996) introduce a new computation method for hue and saturation using log operation.

### 2.1.1.4  YCrCb

$YC_rC_b$ is an encoded and non-linear signal. It has a wide usage in Europe for image compression and communication. Luma component represents brightness and it is constructed as a weighted sum of the RGB values, the other two components namely $C_r$ and $C_b$ are computed by subtracting luma from red and blue channels respectively.

$$Y = 0.299R + 0.578G + 0.114B \qquad \text{(2.3a)}$$

$$C_r = R - Y \qquad \text{(2.3b)}$$

$$C_b = B - Y \qquad \text{(2.3c)}$$

The simplicity of the transformation and the discrimination of luminance and chrominance components make this color space highly popular for skin color modeling (Hsu et al. 2002).

## 2.1.2  Skin Color Modeling

The ultimate goal of skin color modeling is to construct a decision rule to differentiate the skin and non-skin pixels. A common method for this task is introducing a metric, which measures the distance between the color of the current pixel and skin tone model.

### 2.1.2.1   Explicitly defined skin region

One method for building a skin color classifier is to define boundaries explicitly in some color space. For example; Jordao et al. (1999) analyzed statistically large image data sets and concluded that the hue distribution of skin is approximately Gaussian with mean angle $\theta=0^o$ and a variance $\sigma=7^o$. After that conclusion they built a skin color detector by thresholding the hue value of an image using as threshold $T_{low} = \bar{\theta} - 2\sigma$ and $T_{high} = \bar{\theta} + 2\sigma$.

### 2.1.2.2   Nonparametric skin distribution modeling

Nonparametric skin modeling methods estimate skin distribution from the training data without defining an explicit model of skin color. In the literature several face detector algorithms use these kinds of skin modeling such as (Sigal et. al. 2000; Zarit et al. 1999; Chen et al. 1995).

In this skin modeling technique, to classify skin color pixels a histogram based approach is used. The color space, which is generally chrominance plane only, is quantized into a couple of bins corresponds to particular range of color. Each bin holds the count of this individual color components frequency in the training skin images and it is called as look up table (LUT). After the training part, the histogram values in the LUT constitute the likelihood that corresponding colors will correspond to skin (Vezhnevets et al. 2003).

### 2.1.2.3 Parametric skin distribution modeling

Due to the fact that non-parametric skin modeling techniques require much storage space and significantly depend upon the training set's representations, more compact skin model representations are proposed.

One of the parametric skin color distribution modeling technique is Gaussian modeling. In this method, skin color distribution is modeled by an elliptical Gaussian joint probability density function. The function is defined as;

$$p(c|skin) = \frac{1}{2\pi|\Sigma_s|^{1/2}} e^{-\frac{1}{2}(c-\mu_s)^{1/2}\Sigma_s^{-1}(c-\mu_s)} \tag{2.4}$$

In this equation, c is a color vector, $\mu_s$ and $\Sigma_s$ are mean vector and covariance matrix respectively. These parameters are estimated from the training data;

$$\mu_s = \frac{1}{n}\sum_{j=1}^{n} c_j \tag{2.5a}$$

$$\Sigma_s = \frac{1}{n-1}\sum_{j=1}^{n}(c_j - \mu_s)(c_j - \mu_s)^T \tag{2.5b}$$

where n is the total count of skin color samples. Menser and Wien (2000) stated that p(c|skin) probability can be used as a measure of how much c is "skin-like" color.

Another parametric skin color distribution modeling is elliptical boundary model. After several experiments, Lee and You (2002) concluded that the best way to express skin color cluster's shape distribution's model is elliptical. They also suggested that usage of symmetric Gaussian model leads high false positive rates due to asymmetry of skin color cluster along to its density peak. The proposed method namely "elliptical boundary model" is equally fast and simple in the both training and testing part. The model is described as;

$$\Phi(c) = (c - \phi)^T \Lambda^{-1}(c - \phi) \tag{2.6}$$

The first stage of training procedure is eliminating color samples with low frequency to reduce noise and the second stage of the training procedure is estimating the model parameters.

$$\phi = \frac{1}{n}\sum_{i=1}^{n} c_i \tag{2.7a}$$

$$\Lambda = \frac{1}{N}\sum_{i=1}^{n} f_i(c_i - \mu)(c_i + \mu)^T \tag{2.7b}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{n} f_i c_i \tag{2.7c}$$

$$N = \sum_{i=1}^{n} f_i \tag{2.7d}$$

n is the total number of training color vectors of the training skin pixel set, $f_i$ is the number of skin samples of vector $c_i$. Given pixel value with color c is classified as skin when $\Phi(c) < \theta$, where $\theta$ is the threshold value.

## 2.2 COMPARED SKIN COLOR DETECTION METHODS

During the last decades many researchers all over the world tried to solve the skin color segmentation problem and proposed many algorithms. In order to use its benefits in the application of face detection and facial emotion recognition, we first reviewed the task of pixel level skin color segmentation algorithms and assessed their performances. When choosing the algorithms for evaluation, we considered their application areas. Due to the fact that our aim is face detection and facial emotion recognition, we investigated the algorithms that were used in the field of face related applications before.

### 2.2.1 Explicit Method 1

Experimental results show that skin color regions consistently contain a significant level of red color (Ogihara et al. 1996). Building on this observation, one method is to define explicit boundaries to skin cluster in the same color space (Solina et al. 2003).

(R, G, B) is classified as skin if;

$$R > 95 \; and \; G > 40 \; and \; B > 20 \qquad\qquad \textbf{(2.8a)}$$
$$\max\{R, G, B\} - \min\{R, G, B\} > 15 \qquad\qquad \textbf{(2.8b)}$$
$$|R - G| > 15 \; and \; R > G \; and \; R > B \qquad\qquad \textbf{(2.8c)}$$

This method is computationally fast and simple. Therefore, it has been used by many researchers such as Ahlberg (1999), Fleck et al. (1996). The main advantage of this method is its simplicity of skin color detection rule that leads to development of very rapid classifiers.

### 2.2.2 Explicit Method 2

One another method that uses basic thresholding to detect skin color pixels in an image is described in (Haj et al. 2009). In order to model skin color, around 800000 skin pixels were sampled from 64 different images of people with different ethnicities and under various illumination conditions. Then, using the distribution of these pixels, it is concluded that a pixel is skin if;

$$R > 75 \; and \; 20 < R - G < 90 \; and \; \frac{R}{G} < 2.5 \qquad\qquad \textbf{(2.9)}$$

The advantage of this method is also its simplicity and efficiency of computation since there is no need for color space transformation.

### 2.2.3 Skin Probability Map And Bayesian Classifier

Previously mentioned skin color segmentation algorithms differentiate skin pixels from non-skin pixels based on a thresholding method without using any color space transformation. Another solution for classification of skin color pixels is to model the color space using a non-parametric method. In this approach, the distribution of skin color values is estimated from a skin color database and then, to segment a particular

skin pixel value, the probability of being a skin pixel is calculated according to Bayesian probability rule.

Unlike the previous two methods, this method uses a mathematical calculation to classify a pixel as a skin or non-skin. In order to discriminate skin and non-skin colors, the color space is quantized into a number of bins where each of them corresponds a specific range of color components value pairs or triads depends upon the 2D or 3D case. These bins are called as look up table (LUT) and store the particular pixel values' frequency in the training set.

In this task to construct the histogram bins for skin detection, we used the well known and publicly available labeled dataset which is the Compaq dataset (John and Rehg 2002). In the training part, we divided the labeled image dataset into two parts, one for training and one for testing. The training part of the skin dataset includes 3098 images and the non-skin dataset includes 2000 images.

Skin and non-skin color histogram are produced due to the need for calculation of conditional probabilities. Each of them is built as the size of histogram various from 256 bins per channel to 16. Then, the probabilities of observing a given pixel given that it is a skin pixel and non-skin pixel are calculated as;

$$P(rgb|skin) = \frac{s[rgb]}{T_s} \qquad \textbf{(2.10a)}$$

$$P(rgb|{\sim}skin) = \frac{n[rgb]}{T_n} \qquad \textbf{(2.10b)}$$

where s[rgb] is the pixel count in RGB bin of the skin color histogram, n[rgb] is the equivalent count from the non-skin color histogram, and $T_s$ and $T_n$ are the total count of pixel values in the skin and non-skin color histograms respectively.

After that, skin color classifier is derived through the standard Bayes' rule.

$$P(skin|rgb) = \frac{P(rgb|skin)P(skin)}{P(rgb|skin)P(skin)+P(rgb|{\sim}skin)P(\sim skin)} \qquad \textbf{(2.11)}$$

P(rgb|skin) and P(rgb|~skin) are directly computed from skin and non-skin color histograms. The priori probabilities P(skin) and P(~skin) can also be estimated from the overall number of skin and non-skin samples in the training set, $P(skin) = T_n/T_s + T_n$ and likewise $P(\sim skin) = T_s/T_s + T_n$ (John and Rehg 2000). Finally an inequality is used to estimate how much a pixel is "skin-like", $P(skin|rgb) \geq \theta$ where $\theta$ is the threshold value.

## 2.2.4 Classification using YCrCb Color Space

Accuracy of a skin color classifier depends upon many factors and the most important one is illumination. Illumination can be discriminated by human eyes without any effort but it is a challenging task for a computer vision system. Hence, many researchers have tried to tackle the illumination effect onto images. One method to make the algorithm illumination free is to use different color space that is affected by illumination less than naïve RGB color space.

Modeling skin color for segmentation requires a proper color space choice and identification of a cluster for skin pixels in that color space. It has been observed that $YC_rC_b$ color space is relatively less affected by illumination due to the perceptual uniformity and it has been assumed that the chrominance components of the skin color are independent of luminance component (Jain et al 2002). However; in practice, the skin color is non-linearly dependent upon luminance (Jain et al 2002). In order to show the luma dependency of skin color tone, they collected skin color images from Heinrich-Hertz-Institute (HHI) skin color database, and then depicted in different color spaces as seen in Figure 2.1.

As it can be seen from Figure 2.1, the skin pixel values shrink while luma value is increasing and decreasing in the $YC_rC_b$ histogram. Likewise, skin color detection without using luma components results many false positives (Jain et al. 2002). Therefore, skin color is relatively dependent on luma component and removing luma component will not give the expected accuracy.

**Figure 2.1: Training set histograms in YCrCb color space**
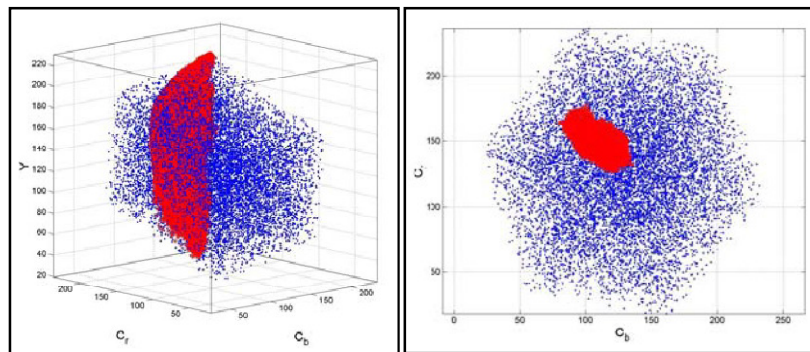


*Source*: Jain et al. (2002)

To get rid of the dependency of luma component, a non-linear transformation has been done to make the skin cluster luma independent in the $YC_rC_b$ color space. The details of this transformation are described in Appendix 1. The transformed training set histograms are shown in Figure 2.2.

**Figure 2.2: Transformed training set histograms in YCrCb color space**
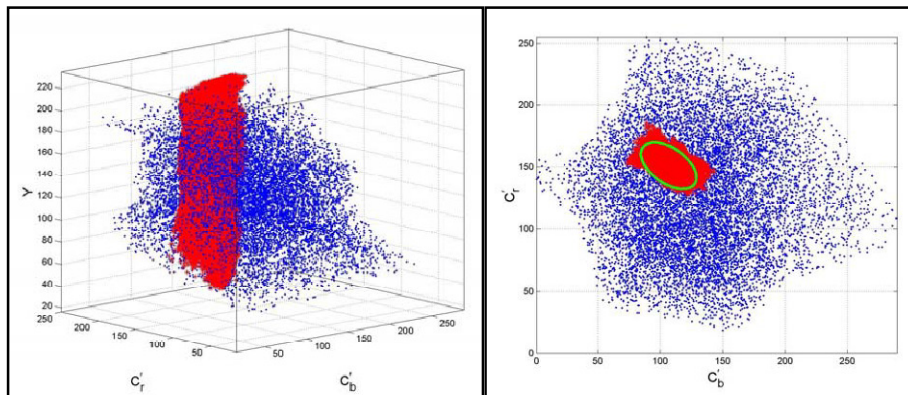


*Source*: Jain et al. (2002)

The transformation makes the skin cluster luma independent, and enables a robust detection of various illumination conditions.

## 2.3  EXPERIMENTAL RESULTS

In order to evaluate the performances of the four skin color segmentation algorithms, we used two datasets. The first dataset is the Compaq skin color dataset. This dataset is freely available for academic purposes and can be downloaded from the web page of (John and Rehg 2002). The Compaq dataset consists of three parts, namely images containing skin colored regions, the corresponding manually labeled images and images that do not contain skin colored regions. We used this dataset to estimate performances of the analyzed skin color detection algorithms since the ground truth comes from the dataset. The other dataset is Faces in the Wild (FDDB). This dataset was constructed for mainly the aim of evaluation of different face detection algorithms. Even though the exact face locations are given in this dataset, there is no ground truth for skin colored pixels. However, it is a good dataset to evaluate a skin color detection algorithm visually.

In Table 2.1, we compare the skin color detection performances of four methods denoted with the following acronyms; TP, FP and MR. They denote the number of true positives, false negative and miss rates, respectively. The precision is defined as $Precision = TP/(TP + FN)$. We can observe from the Table 2.1 that the highest recall is achieved by the $YC_rC_b$ based classifier.

Although method 1 and method 2 are simple and fast, as we can see from the Table 2.1 these methods gave significantly good results and in some situations they outperformed the other segmentation methods. However, yellow like colors could not be eliminated using these methods in a few images Figure 2.3. Unlike the previous two methods, Bayesian skin color classifier is not that simple and has a little computational load. Nevertheless, when we look at the Bayesian classifier's true positive rate, this segmentation method gave significantly good results. The last method namely, $YC_rC_b$ , against the long odds, gave the least true positive rate but its false positive rates is the best amongst the other classifiers that are evaluated. According to the Table 2.1, the results show that method 2 and Bayesian methods give the best results according to the true positive rate and these two methods will be used later in a face detector as a pre-step.

**Table 2.1: The results of skin color classifiers on Compaq dataset**

| Skin Color Classifier | TP Rate | FP Rate | Miss Rate | Precision |
|---|---|---|---|---|
| Method I | 78.00% | 18.21% | 21.99% | 81.78% |
| Method II | 87.85% | 19.61% | 12.14% | 80.38% |
| Skin Probabilit Map And Bayesian C. | 86.89% | 17.65% | 13.10% | 82.34% |
| $YC_rC_b$ | 46.40% | 13.01% | 53.59% | 86.98% |

**Figure 2.3: Original image, method 1 result, method 2 result**



Despite the numerical results, we also show the results visually in Figure 2.4 and it can be seen from the images how the methods work properly. Briefly, method 1 (RGB based method) can detect the skin regions accurately but in some cases it holds unnecessary regions such as eyebrows, lips etc.; on the other hand, method 2 (RGB based method) and Bayesian classifier are relatively yielded better results. As we can see from the images, they mostly eliminate unrelated regions. The last method which is $YC_rC_b$ eliminates unrelated regions very accurately but it cannot detect shadow like skin color regions, therefore it produces very low true detection rate. In first image of the Figure 2.4, woman's leg cannot be separated much and in the second image tennis player's arm cannot be segmented very accurately either.

**Figure 2.4: A few results of skin color classifiers on the FDDB. From left to right the original image and the results of the method 1, method 2, Bayesian classifier and YCrCb based segmentation, respectively**

# 3. FACE DETECTION USING MOMENT INVARIANTS

Face detection is one of the most studied problems in the history of computer vision; nevertheless researches are still struggling to solve the face localization problem more accurately. The need for face detection comes with the increasing technology and rapid increase of computational power. Many products, regardless of academic or commercial purpose, have the capability for a computer to interact with humans by looking at people through cameras, listening people through microphones. Then these kinds of inputs are processed by an intelligence system, which respond people in a friendly manner called human-computer interaction (HCI).

Without any doubt, face detection is one of the most used techniques that enable an interaction between computers and humans; hence, face detection in still images or video sequences is a preliminary step for many facial analysis algorithms such as face tracking, face recognition, face modeling, facial expression recognition, age/gender recognition etc. The aim of face detection is to determine any faces in the image, if present, and return the image location with a correct size. This is a very challenging task for a computer vision system, while it is a trivial task for a human being. The difficulty in a face detection system is to localize a face under many variations of scale, location, orientation, pose, facial expression, illumination, occlusion etc. Therefore, face detection is an expensive search problem.

## 3.1  LITERATURE SURVEY

Many algorithms have been reported about face detection and are surveyed in (Yang et al. 2002; Hjelmas et al. 2001). For example, Yang et al. (2002) grouped the presented methods into four categories; knowledge based methods, feature based methods, template matching methods and appearance based methods (Zhang and Zhang 2010). Briefly, knowledge based methods use pre-defined rules based on human knowledge to determine face locations, feature based methods find faces in an image using robust features under different pose and illumination conditions, template based methods use pre-defined face templates to judge if a region is whether a face or not, appearance

based methods learn face models using representativeness of faces in a training set to perform face detection.

There have been many contributions in the field of face detection during the past decades. The algorithms in the literature mainly implement the face detection task as a binary pattern classification task. That is, the content of a specific region of an image is transform into robust features and then a classifier that is trained on example faces decides whether the region is a face or not. However, here below, we discuss the general approaches for face detection in the literature.

### 3.1.1 Knowledge Based Methods

In this approach, the rules for developing a face detector are based upon researcher's knowledge of human faces (Kotropoulos and Pitas 1997; Yang and Huang 1994; Kanade 1973). It is easy to derive a simple rule to describe facial features and their relationships. For instance, a face often appears in an image with two eyes that are symmetric to each other, a nose and a mouth. The relationships between facial features can be given by their relative distances and positions. In this approach, firstly, facial features are extracted from an input image then face candidate regions are identified based upon the coded rules. Sometimes a verification step is applied to reduce to false detection rate.

One problem occurs when translating human knowledge into a well-defined rule chain. If the rules are strict, they might fail to detect faces since it is hard to pass all the strict rules, on the other hand, if the rules are too general, they might give many false detected faces, and extending this approach to all possible cases such as rotated faces is difficult.

Yang and Huang (1994) used a hierarchical knowledge-based algorithm for face detection. The algorithm consists of three levels. At the highest level, they estimate the whole possible face candidate regions by scanning a window over the input image while applying a set of rules. The rules at this level have only general descriptions what a face looks like and while the level are increasing, the rules become more detailed. At the second level, after local histogram equalization is performed, edge detection is applied,

and then survived candidate regions are passed through a final set of rules. If a region passes all the levels, it is accepted as a face.

### 3.1.2 Feature Based Methods

Unlike the knowledge based methods, researchers have been trying to find invariant features of faces for face detection task. Due to the fact that human can detect faces and objects in different poses and lighting conditions without any effort, researchers suggest that there must exist some properties or features which are invariant over these variations. Many methods have been proposed to extract facial features, and judge the presence of a face (Sirohey 1993; Chetverikov and Lerch 1993). Facial features such as mouth, nose, eyes, eye brows etc. are extracted using well-known edge detectors. Statistical models are built to describe the relationship between extracted features and verify the presence of a face. One problem in these feature based algorithms is that feature boundaries might be weak for faces, and shadows can cause many strong edges which make the algorithm useless.

Leung et al. (1995) proposed a probabilistic method to localize a face in an image based upon local feature detection and random graph matching. They suggested the face detection task as a search problem in which the aim is to find a face-like pattern using certain facial features. Five features are used to describe a face, namely two eyes, two nostrils and nose/lip junction. They used the relative distances between features to judge whether a region is a face or not since the facial features cannot appear in arbitrary arrangements. The expected locations of each feature are estimated using a statistical model over a number of faces in a data set.

### 3.1.3 Template Based Methods

In template based methods, a standard face pattern, generally frontal is defined and parameterized by a function manually (Govindaraju et al. 1990; Craw et al. 1987). In a test image, the correlations with the standard patterns are computed for each region of a face independently. The presence of a face is determined based upon the correlation values. The main advantage of these kinds of methods is to be simple and easy to

implement. However, dealing with variation in scale, pose and shape makes these algorithms inadequate in some situations. Example face template can be seen in Figure 3.1. This template is composed of 16 regions (the gray boxes) and 23 relations (shown by arrows).

**Figure 3.1: A 14x16 pixel ratio for face localization.**
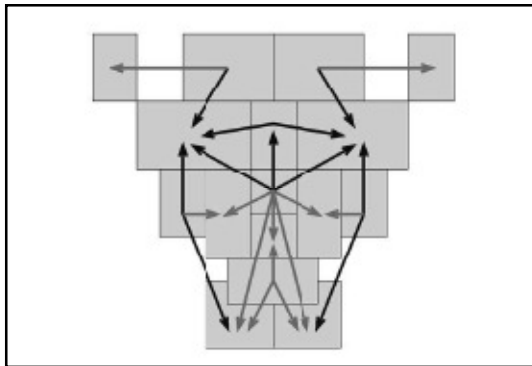


*Source*: Ahuja et al. (2002)

Sinha et al. (1994)'s work can be given as an example of this method. He described the space of face patterns using a set of spatial image invariants. His idea based upon the fact that, while the illumination changes particular parts of faces, the relative brightness of these particular parts remains unchanged. If the brightness constraints between facial parts are satisfied, the region is accepted as a face.

### 3.1.4 Appearance Based Methods

In this method, unlike the template based method the "templates" are learnt from example images (Rowley et al. 1998; Osuna et al. 1997; Turk and Pentland 1991). Appearance based methods generally rely on statistical or machine learning techniques to find relevant characteristics of a face or non-face.

Appearance based methods can be understood in a probabilistic framework. A feature vector derived from an image is evaluated as a random variable x, and faces or non-faces characterize this random variable using class-conditional density functions

$p(x|face)$ and $p(x|non-face)$. One of the well-know classifier can be used to classify candidate region as face or non-face.

Another approach is to find a discriminant function between face and non-face classes. Recently, support vector machines have been proposed for this task. In these methods, facial features are projected to a different dimensional space to form a decision surface between the projected face and non-face patterns.

### 3.1.5 Viola Jones Face Detector

Amongst the face detectors the one based upon a learning algorithm has attracted much attention Viola Jones (2001) and it gives very good results. This face detector has been developed by Viola and Jones (2001) and it can be accepted as a feasible face detector for real world applications.

The structure of the Viola Jones face detector based upon three main ideas which make it a successful face detector. The algorithm can run in real time; and uses the concept of the integral image, learning via *AdaBoost* and a cascade structure.

### 3.1.5.1 Integral image

Integral image, which is also known as a summed area table, is an algorithm to compute the sum of values in a rectangle that has been proposed by Crow (1984). Viola-Jones then applied the integral images for extraction of Haar-like features, as detailed below. Given an image $i(x, y)$, the integral image $ii(x, y)$ is calculated as follows;

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y') \tag{3.1}$$

The computation the sum pixels in a rectangle is extremely efficient using this method as shown in Figure 3.1.

**Figure 3.2: Illustration of the integral image**

**and Haar-like features (a-f)**
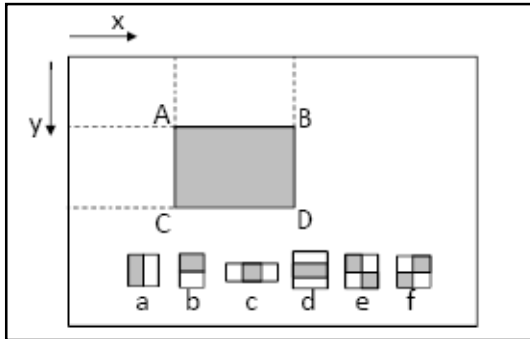


*Source*: Zhang and Zhang (2010)

The sum of pixel intensities in the region ABCD can be calculated as;

$$\sum_{(x,y)\in ABCD} i(x,y) = ii(D) + ii(A) - ii(B) - ii(C) \qquad \textbf{(3.2)}$$

which requires only four array references.

Simple Haar-like features can be computed using the integral image as shown in Figure 3.2 (a-f). The features are defined as the weighted intensity difference between sub-rectangles. For example, the feature value is calculated using difference of the gray and white rectangles in feature (a).

### 3.1.5.2 AdaBoost learning

Boosting is a method of finding an accurate hypothesis by combining many weak hypotheses, which have moderate accuracy. Details can be found in (Meir and Ratsch 2003; Friedman et al. 1998).

Adaptive Boosting algorithm is often considered as a pre-processing for more practical boosting algorithms. There is a generalized version of *AdaBoost*, referred to as *RealBoost* and some the researchers (Wu et al. 2004; Bishop and Viola 2003) state that *RealBoost* algorithm yields better results than the original *AdaBoost.*

Let consider a set of training examples, $S = \{(x_i, z_i), i = 1, ...., N\}$ where $x_i$ belongs to an instance space $X$ and $z_i$ belongs to a finite label space $Z$. For a binary classification problem, $Z = \{1, -1\}$, where $z_i = -1$ for negative samples and $z_i = 1$ for positive samples. *AdaBoost* produces an additive model to predict the label of an input sample $x$. From the statistical point of view, *AdaBoost* fits an additive logistic regression model using expectation minimization.

### 3.1.5.3 Cascade structure

Cascade classifier is one of the key components in the Viola Jones detector. It is small and thus more efficient, while keeping almost all positive examples it rejects most of the negative sub-windows.

The process of the classifier begins with a decision tree, which is called *cascade* (Viola and Jones 2001). As shown in Figure 3.3, input sub-windows pass through a set of nodes during detection stage. Each of the nodes makes a binary decision whether the sub-window will be kept or not. Consequently, if the sub-window passes all the nodes, it is accepted as a face.

**Figure 3.3: Cascade classifier**
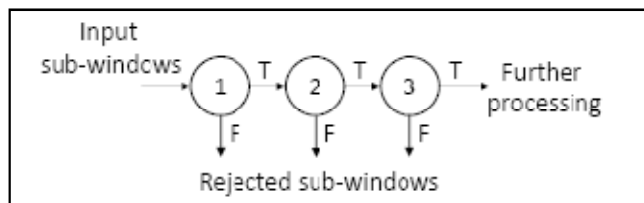


*Source*: Zhang and Zhang (2010)

The Viola Jones face detection algorithm has the ability of detecting faces under the presence of up to ±15 degree in plane and up to ±45 degree out of plane rotation (Viola Jones 2004) since as it can be seen from Figure 3.2, Haar-like features are horizontal, vertical or diagonal. However, when the rotation of the faces exceeds the given angles, the algorithm fails. On the other hands, the VJ algorithm uses only brightness information in a search window, resulting in a high false positive rate due to the face-like brightness patterns in the background.

## 3.2   PROPOSED FACE DETECTOR

Frontal face detection task has been studied very well. On the other hand, multi view face detection, however, is still an open topic. Although some attempts were made to detect multi view faces (Huang et al. 2004), it can be accepted as a challenging problem due to fact that methods are either computationally expensive or they produce a large number of false positives (Wang and Ji 2004), or both.

In this thesis we propose a multi view face detection algorithm for color images in the presence of facial expression. The methods starts by detection of skin blobs in a given image, and then a small number of features are extracted from these blobs. Finally, these features are used to learn a statistical model of faces or non-faces.

### 3.2.1 Extraction Of Facial Features

After skin color segmentation using one of methods described in section 2, every connected skin color region is treated as a face candidate. However, to understand whether the skin color region is a face or not, a few robust features will be extracted from these regions, and then the features will be used to train several classifiers.

For training the classifiers the CVL database (Haj et al. 2009; Peer et al. 2003) is used, which contains color images of people under seven different poses with variations in the expression and contains a total 979 face images. Some example images from CVL database can be seen Figure 3.4. CVL database is used as the positive training set and is used to learn the face features. For negative training set, many images that do not contain any faces but contain skin color pixels (skin color regions) have been selected from the Compaq (John and Rehg 2002) dataset and from several other sources.

**Figure 3.4: Some sample images from the training set, the CVL Database**
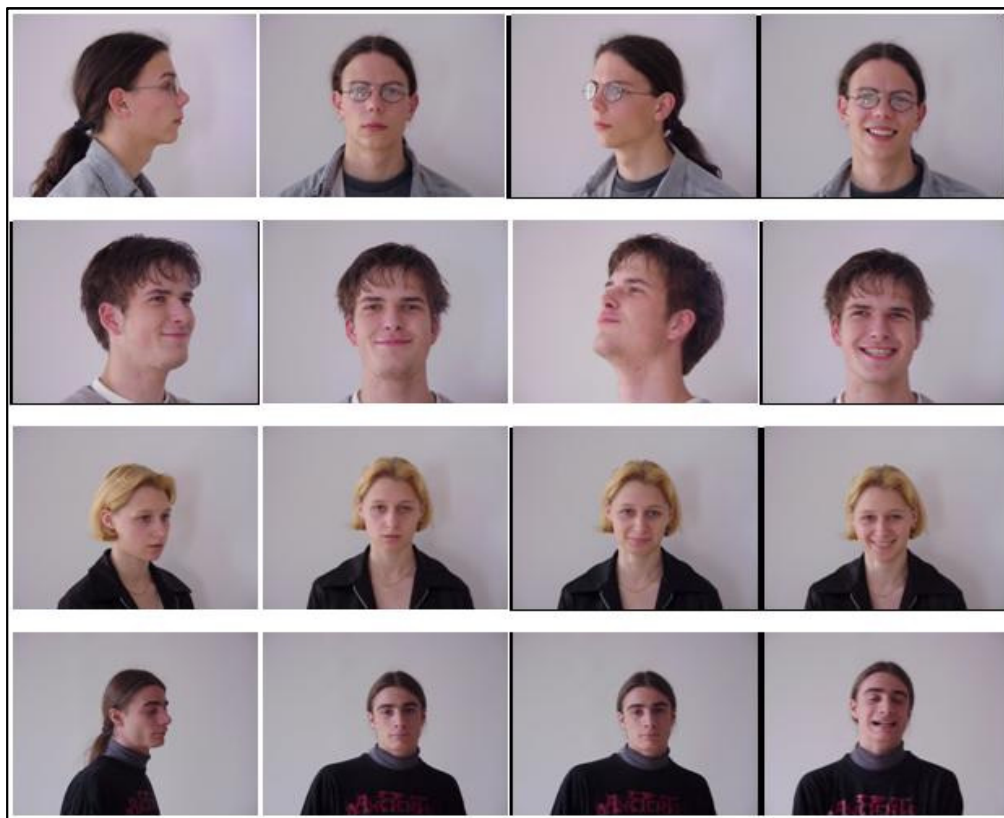


*Source*: CVL database (Peer et al. 2003)

In order to extract robust facial features, the face regions should be distinguished as much as possible from the other parts of the image, therefore, after the skin color segmentation process, a bounding rectangle is fit to each segmented region. These rectangles are then used to extract features. In some cases, these rectangles may have some part of other skin blobs. In order to eliminate those parts, only the biggest connected skin color region is kept since we know that the biggest skin color part of an image in the CVL dataset is a face Figure 3.5.

**Figure 3.5: Example binary images**



*Source*: CVL dataset

The extraction of these rectangles is the key point for the proposed classifier. They are used in two ways. First, they are used for extracting geometrical shape features from the binary images and second, they are used for extracting moment features from the image texture in the original image. Both features are very robust to geometric variations of faces and they are elaborated below.

**3.2.2 Geometric Shape Features**

In general, a shape descriptor is a set of numbers to describe a given shape feature. A descriptor should quantify the shape considering human intuitions or task-specific requirements and should meet the following requirements (Mingqiang et al. 2008);

    i.   the descriptor should represent the content of the information as much as possible

   ii.   the descriptor vector's size must not be too large

  iii.   the distance between descriptors should be computed rapidly

Considering the shape feature specifications, we choose 9 geometric features to describe a face blob, namely aspect ratio, eccentricity, euler number, extent, orientation, solidity, roundness and centroid position (Haj et al. 2009).

i. Aspect ratio

The aspect ratio of a blob is the ratio of the longer side to the shorter side.

ii. Eccentricity

Eccentricity is a measure of the aspect ratio and it can be thought of as a measure of how much the ellipse (which is fit to the skin blob region) deviates from being circular. It can be calculated like that;

$$\varepsilon = \sqrt{\frac{a^2 - b^2}{a^2}} \tag{3.3}$$

Typical ellipse parameters can be seen in Figure 3.6.

**Figure 3.6: Ellipse parameters**
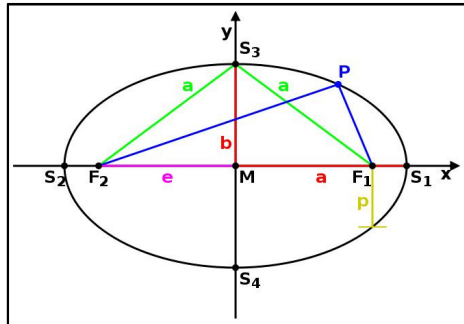


*Source*: Wikipedia

iii. Euler number

Euler number is a useful topological descriptor and is defined as the number of connected components C minus the number of holes H.

iv. Extent

A scalar that specifies the ratio of the number of pixels in the skin blob region to the number of pixels in the total bounding box.

v.  Orientation

Orientation is the angle between the x-axes to the major axis of the ellipse that is fit to the skin blob region.

vi.  Solidity

Solidity describes the extent to which the shape is convex or concave (Cheng et al. 2001).

$$Solidity = A_s/H \tag{3.4}$$

where, $A_s$ is the area of skin color region and $H$ is the convex hull area of the shape. The solidity of a convex shape is always 1 (Mingqiang et al. 2008).

vii.  Roundness

Roundness is defined as the ratio of the minor axes of the ellipse to its major axes.

viii.  Centroid position

The centroid position of an object is the point within that object from which the force of gravity appears to act.

Two features encoding the Euclidian distance between the centroid of the blob and the center of its bounding box in both x and y directions.

### 3.2.3 Moment Invariant Features

In the field of image processing and computer vision, an image moment is a certain particular weighted average of the image pixels' intensities. Moments of images represent a lot of information and it has been widely used for a long time, especially in pattern recognition applications. Their main advantage is their robustness to variations such as rotation, translation and scale. Various kinds of moments have been proposed

for the object recognition task, but amongst them, geometric, R and Zernike moments have been quite popular.

### 3.2.3.1 Geometric moments

Geometric moments, which are also known as Hu moments, are the simplest among the moment functions with basis $\varphi_{pq} = x^p y^q$ (Hu 1962). Geometric moment function $m_{pq}$ is given as;

$$m_{pq} = \sum_x \sum_y x^p y^q f(x, y) \qquad p, q = 0,1,2 \ldots \ldots \ldots \tag{3.5}$$

where $(p + q)$ is the order of the moment function. Corresponding central moments are defined as;

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x, y) \qquad p, q = 0,1,2 \ldots \ldots \ldots \tag{3.6}$$

where $\bar{x} = m_{10}/m_{00}$ and $\bar{y} = m_{01}/m_{00}$ and a set of 7 invariant moments are given by (Hu 1962);

$$
\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + \\
& \quad (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
\phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{03} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + \\
& \quad 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + \\
& \quad (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned}
\tag{3.7}
$$

where $\eta_{pq} = \mu_{pq}/\mu_{00}^\gamma$ and $\gamma = 1 + (p + q)/2$ for $(p + q) = 2,3,4, \ldots$

$$\iint_{\Omega} p_{pq}(x,y)p_{mn}(x,y)dxdy = 0 \qquad\qquad \textbf{(3.8a)}$$

$$\iint_{\Omega} w(x,y)p_{pq}(x,y)p_{mn}(x,y)dxdy = 0 \qquad\qquad \textbf{(3.8b)}$$

Geometric moments are invariant to rotation, scaling and translation; however they are not orthogonal (OG) since the polynomial basis of this moment set does not satisfy the condition of orthogonality (3.8a-3.8b). The main reasons of usage OG basis are; stable and fast numerical implementation (Flusser et al 2009), which could be efficiently implemented and the second reason of usage OG basis is for better image reconstruction. Below, an example demonstrates the computation of invariant moments using (3.7) for various images and compares the results. A black background is added to all images to make them of equal size since zero pixels do not affect the computation of moments. To reduce the dynamic range of moments, they are passed through a logarithmic (i.e. $sgn(\phi_i)log_{10}(|\phi_i|)$) function. Absolute value is used for the negative and fractional results; the *sgn* function is used to preserve the sign.

**Figure 3.7: Test images used for moment computation**
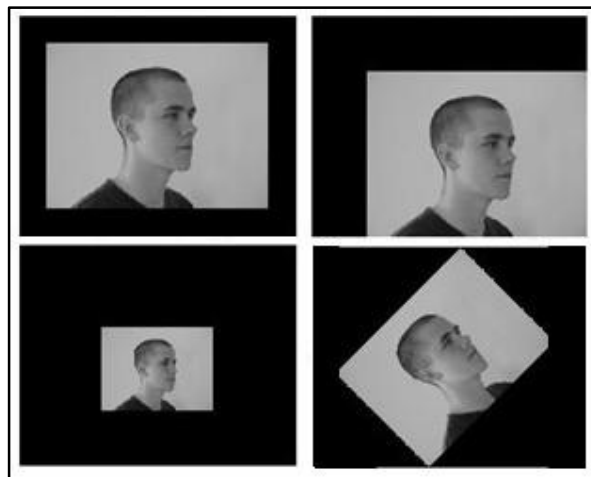


*Source*: CVL dataset

As we can see from Figure 3.7, images are translated, scaled and rotated by $45^o$. However, the moment invariants of each image are almost the same as can be seen clearly from Table 3.1.

**Table 3.1: Geometric moment invariants of the images in Figure 3.7**

| Moment Invariants | Original Image | Translated Image | Scaled Image (50%) | Rotated Image $45^o$ |
|---|---|---|---|---|
| $\phi_1$ | 6.7324 | 6.7324 | 6.7325 | 6.7320 |
| $\phi_2$ | 15.6711 | 15.6711 | 15.6712 | 15.6598 |
| $\phi_3$ | 26.1334 | 26.1334 | 26.1331 | 26.1455 |
| $\phi_4$ | 28.2974 | 28.2974 | 28.2983 | 28.3142 |
| $\phi_5$ | -56.7535 | -56.7535 | -56.7556 | -56.8035 |
| $\phi_6$ | -37.7370 | -37.7370 | -37.7387 | -37.7639 |
| $\phi_7$ | 55.5652 | 55.5652 | 55.5664 | 55.5945 |

### 3.2.3.2 R moments

Liu et al. (2008) has stated that geometric moments are not scale invariant in the discrete case. Therefore, he proposed R moments as an improvement over geometric moments. These R moments are obtained from the geometric moments; as follows:

$$R_1 = \frac{\sqrt{\phi_2}}{\phi_1}, \qquad R_2 = \frac{\phi_1 + \sqrt{\phi_2}}{\phi_1 - \sqrt{\phi_2}}$$

$$R_3 = \frac{\sqrt{\phi_3}}{\sqrt{\phi_4}}, \qquad R_4 = \frac{\sqrt{\phi_3}}{\sqrt{|\phi_5|}}$$

$$R_5 = \frac{\sqrt{\phi_4}}{\sqrt{|\phi_5|}}, \qquad R_6 = \frac{|\phi_6|}{\phi_1 \phi_3} \tag{3.9}$$

$$R_7 = \frac{|\phi_6|}{\phi_1 \sqrt{|\phi_5|}}, \qquad R_8 = \frac{|\phi_6|}{\phi_3 \sqrt{|\phi_2|}}$$

$$R_9 = \frac{|\phi_6|}{\sqrt{\phi_2 |\phi_5|}}, \qquad R_{10} = \frac{|\phi_5|}{\phi_3 \phi_4}$$

**Table 3.2: R Moment invariants of the images in Figure 3.7**

| R Moment Invariants | Original Image | Translated Image | Scaled Image (50%) | Rotated Image 45$^o$ |
|---|---|---|---|---|
| $R_1$ | 0.5880 | 0.5879 | 0.5878 | 0.5878 |
| $R_2$ | 3.8544 | 3.8543 | 3.8522 | 3.8522 |
| $R_3$ | 0.9610 | 0.9609 | 0.9609 | 0.9609 |
| $R_4$ | 0.6785 | 0.6785 | 0.6784 | 0.6784 |
| $R_5$ | 0.7061 | 0.7061 | 0.7060 | 0.7060 |
| $R_6$ | 0.2144 | 0.2144 | 0.2145 | 0.2145 |
| $R_7$ | 0.7440 | 0.7440 | 0.7442 | 0.7442 |
| $R_8$ | 0.3647 | 0.3647 | 0.3649 | 0.3649 |
| $R_9$ | 1.2653 | 1.2654 | 1.2661 | 1.2661 |
| $R_{10}$ | 0.0767 | 0.0767 | 0.0767 | 0.0767 |

As it can be seen from Table 3.2, R moment invariants take the almost same values in various geometric conditions and it makes them a robust feature.

### 3.2.3.3 Zernike moments

Unlike geometric and R moments, Zernike moments are orthogonal (Celebi and Aslandogan 2005). Teague proposed the use of orthogonal moments in (Teague 1980). He stated that Zernike moments do not carry redundant information and their noise sensitivity is very low as opposed to geometric or R moments due to the its orthogonal basis.

Computation of Zernike moments consists of three steps; computation of radial polynomials, computation of basis functions and computation of moment by projecting the image onto basis functions.

Computation of Zernike moments begins with the computation of radial polynomials which is defined as in 1D;

$$R_{nm}(\rho) = \sum_{s=0}^{(n-|m|)/2} c(n,m,s)\rho^{n-2s} \qquad \textbf{(3.10)}$$

and;

$$c(n, m, s) = (-1)^s \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!} \tag{3.11}$$

In (3.10), $n$ and $m$ are the order and repetition respectively. The parameter $n$ should be a non-negative integer number, and the repetition $m$ should also be an integer with constraints: satisfies $n - |m| = (even)$ and $|m| \leq n$. The following orthogonality property is satisfied by radial polynomials for the same repetitions:

$$\int_0^{2\pi} \int_0^1 R_{nm}(\rho, \theta) R_{n'm}(\rho, \theta) \rho d\rho d\theta = \begin{cases} \frac{1}{2(n+1)} & if \ n = n' \\ 0 & otherwise \end{cases} \tag{3.12}$$

Using the radial polynomials 2D Zernike basis function within a unit circle are formed by;

$$V_{nm}(\rho, \theta) = R_{nm}(\rho) \exp(jm\theta), \quad |\rho| \leq 1 \tag{3.13}$$

Zernike basis functions are orthogonal and satisfy;

$$\int_0^{2\pi} \int_0^1 V_{nm}^*(\rho, \theta) V_{pq}(\rho, \theta) \rho d\rho d\theta = \begin{cases} \frac{\pi}{n+1} & if \ n = p, m = q \\ 0 & otherwise \end{cases} \tag{3.14}$$

Due to the orthogonality property, information between orders and repetitions has no redundancy and overlap.

Finally, complex Zernike moments of order n and with repetition m are defined as;

$$Z_{nm} = \frac{n+1}{\pi} \int_0^{2\pi} \int_0^1 f(\rho, \theta) V_{nm}^*(\rho, \theta) \rho d\rho d\theta \tag{3.15}$$

where in (3.15), $f(\rho, \theta)$ is the image function and '*' denotes the complex conjugate.

As it can be seen from (3.15), Zernike moments are calculated using the inner product of the image and the Zernike basis functions.

In the discrete case, integral operators in the (3.15) are transformed to summation operators and calculation of Zernike moments for digital images can be defined as;

$$Z_{nm} = \frac{n+1}{\lambda_N} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x,y) R_{nm}(\rho_{xy}) \exp(-jm\theta_{xy}) \qquad \textbf{(3.16)}$$

where $0 \leq \rho_{xy} \leq 1$ and $\lambda_n$ is the number of pixels located in the unit circle by mapping transform and it is called the normalization factor.

Zernike moments are widely used in the field of computer vision because of the following advantages.

    i.      The magnitudes of Zernike moments are invariant to rotation (Khotanzad and Hong 1990).

    ii.     They are robust the noise and other minor variations in shape (Kim and Kim 1998).

    iii.    Information redundancy is significantly minor due to the orthogonal basis (Teh and Chin 1988).

### 3.2.3.4  Flusser moments

Flusser stated that some of the invariant moments miss some information of images due to several reasons and he proposed a new moment set called Flusser moments (Flusser et al. 2009). Unlike the other invariant moments, this moment set does not have any dependency to each other.

In order to compute the Flusser moments, first a set of complex moments needed to be computed. Complex moment set can be constructed using the formula below;

$$c_{pq} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x+iy)^p (x-iy)^q f(x,y) dx dy \qquad \textbf{(3.17)}$$

Then, a set of rotation invariants $\mathcal{B}$ is constructed, using the following formula;

$$\mathcal{B} = \left\{ \Phi(p,q) \equiv c_{pq} c_{p_0 q_0}^{p-q} \mid p \geq q \wedge p + q \leq r \right\} \tag{3.18}$$

which are used to calculate moment:

$$\Phi(p_0, q_0) = \frac{\prod_{i=1}^{n_1} \Phi(p_i, q_i)^{k_i} \prod_{i=n_1+1}^{n} \Phi^*(p_i, q_i)^{k_i}}{\prod_{i=1}^{m_1} \Phi(s_i, t_i)^{l_i} \prod_{i=m_1+1}^{m} \Phi^*(s_i, t_i)^{l_i}} \tag{3.19}$$

Substituting (3.18) into (3.19) and grouping the factors, we obtain:

$$\Phi(p_0, q_0) = \frac{c_{p_0 q_0}^{\Sigma_{i=1}^{n_1} k_i(p_i - q_i)} c_{p_0 q_0}^{\Sigma_{i=n_1+1}^{n} k_i(q_i - p_i)} \prod_{i=1}^{n_1} c_{p_i q_i}^{k_i} \prod_{i=n_1+1}^{n} c_{q_i p_i}^{k_i}}{c_{p_0 q_0}^{\Sigma_{i=1}^{m_1} l_i(s_i - t_i)} c_{p_0 q_0}^{\Sigma_{i=m_1+1}^{m} l_i(s_i - t_i)} \prod_{i=1}^{m_1} c_{s_i t_i}^{l_i} \prod_{i=m_1+1}^{m} c_{t_i s_i}^{l_i}} \tag{3.20}$$

Then, 6 moment invariant features such as $\Phi(1,1)$, $\Phi(2,1)$, $\mathrm{Re}(\Phi(2,0))$, $\mathrm{Im}(\Phi(2,0))$, $\mathrm{Re}(\Phi(3,0))$, $\mathrm{Im}(\Phi(3,0))$ are calculated from (3.20).

Flusser gives an example to show the accuracy of the new moment set in his book (Flusser et al. 2009). He used two synthetic images to extract the invariant features namely $f(x, y)$ and $g(x, y)$. In the case of synthetic data, such images $g(x, y)$ exists for any given $f(x, y)$ and can be designed using the following formula;

$$g(x, y) = (f(x, y) + f(-x, y) + f(x, -y) - f(-x, -y))/2 \tag{3.21}$$

Two synthetic image examples are shown in Figure 3.8.

**Figure 3.8: Example synthetic images**
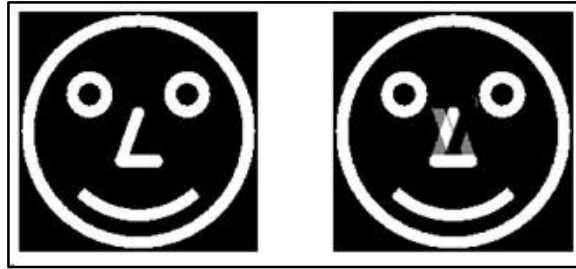


*Source*: Flusser et al. (2009)

**Table 3.3: Flusser and geometric moments of the synthetic images**

| Geometric Moments | Synthetic Image 1 | Synthetic Image 2 | Flusser Moments | Synthetic Image 1 | Synthetic Image 2 |
|---|---|---|---|---|---|
| $\phi_1$ | 1.1843 | 1.1843 | $\Phi(1,1)$ | 1.6809 | 1.6809 |
| $\phi_2$ | 3.7719 | 3.7718 | $\Phi(2,1)$ | 1.2004 | 1.2004 |
| $\phi_3$ | 4.6469 | 4.6469 | $\mathbf{Re(\Phi(2,0))}$ | **2.4811** | **-3.2371** |
| $\phi_4$ | 5.0556 | 5.0558 | $Im(\Phi(2,0))$ | 5.7135 | 5.7135 |
| $\phi_5$ | -10.0000 | -9.9995 | $Re(\Phi(3,0))$ | 7.0818 | 7.0818 |
| $\phi_6$ | 6.9667 | 6.9696 | $Im(\Phi(3,0))$ | -2.0435 | -2.0435 |
| $\phi_7$ | 9.9988 | 10.0000 | | | |

**Table 3.4: R moments of the synthetic images**

| R Moments | Synthetic Image 1 | Synthetic Image 2 |
|---|---|---|
| $R_1$ | 1.6398 | 1.6398 |
| $R_2$ | -4.1255 | -4.1256 |
| $R_3$ | 0.9587 | 0.9587 |
| $R_4$ | 0.6816 | 0.6817 |
| $R_5$ | 0.7110 | 0.7110 |
| $R_6$ | 1.2658 | 1.2664 |
| $R_7$ | 1.8602 | 1.8610 |
| $R_8$ | 0.7719 | 0.7722 |
| $R_9$ | 1.1343 | 1.1348 |
| $R_{10}$ | 0.4256 | 0.4256 |

It can be seen from the Table 3.3, Table 3.4 and Table 3.5 that the R moments for both images are same. The only difference between the moments of two synthetic images is

for Im($\Phi(2,0)$). Flusser moments have the ability to distinguish these kinds of images unlike the other moments.

**Table 3.5: Zernike moments of the synthetic images**

| Zernike Moments | Synthetic Image 1 | Synthetic Image 2 |
|---|---|---|
| $Z_1$ | 0.3183 | 0.3183 |
| $Z_2$ | 0.0000 | 0.0000 |
| $Z_3$ | 0.2691 | 0.2691 |
| $Z_4$ | 0.0188 | 0.0188 |
| $Z_5$ | 0.0479 | 0.0479 |
| $Z_6$ | 0.0441 | 0.0441 |
| $Z_7$ | 0.2453 | 0.2453 |
| $Z_8$ | 0.0400 | 0.0401 |
| $Z_9$ | 0.0229 | 0.0229 |
| $Z_{10}$ | 0.1753 | 0.1754 |
| $Z_{11}$ | 0.1320 | 0.1319 |
| $Z_{12}$ | 0.0021 | 0.0021 |
| $Z_{13}$ | 0.0693 | 0.0693 |
| $Z_{14}$ | 0.0089 | 0.0080 |
| $Z_{15}$ | 0.0781 | 0.0781 |
| $Z_{16}$ | 0.0012 | 0.0012 |
| $Z_{17}$ | 0.1490 | 0.1488 |
| $Z_{18}$ | 0.1433 | 0.1432 |
| $Z_{19}$ | 0.0066 | 0.0066 |
| $Z_{20}$ | 0.0077 | 0.0077 |
| $Z_{21}$ | 0.2888 | 0.2888 |
| $Z_{22}$ | 0.0735 | 0.0757 |
| $Z_{23}$ | 0.1169 | 0.1169 |
| $Z_{24}$ | 0.0002 | 0.0002 |
| $Z_{25}$ | 0.0012 | 0.0012 |
| $Z_{26}$ | 0.0779 | 0.0772 |
| $Z_{27}$ | 0.0342 | 0.0343 |
| $Z_{28}$ | 0.0016 | 0.0016 |
| $Z_{29}$ | 0.0382 | 0.0382 |
| $Z_{30}$ | 0.0017 | 0.0017 |
| $Z_{31}$ | 0.6230 | 0.6229 |
| $Z_{32}$ | 0.0946 | 0.0934 |
| $Z_{33}$ | 0.0781 | 0.0781 |
| $Z_{34}$ | 0.0072 | 0.0072 |
| $Z_{35}$ | 0.0124 | 0.0124 |
| $Z_{36}$ | 0.0001 | 0.0001 |

**3.2.4 Classifiers**

In order to classify the extracted features, we utilized various well-known classifiers: linear Bayes, quadratic Bayes, decision tree, k-nearest neighbor and support vector machines (SVM).

1. Linear Bayes classifier (LB)

   The Bayes classifier is based upon Bayesian probability theorem and despite its simplicity; it can often outperform other complex classification methods when the dimensionality of the features is high. Linear Bayes classifier classifies the inputs into two classes with a linear decision boundary, assuming normally distributed classes and equal covariance matrices.

2. Quadratic Bayes classifier (QB)

   The quadratic Bayes classifier uses Bayesian probability theorem like linear Bayes; however, it tries to separate classes of objects by a quadratic surface.

3. Decision tree (DT)

   Decision trees present rules in contrast to other classifiers. Rules can  be expressed so that human can understand them. In most general terms, building a decision three is to determine a set of if-then-else conditions to allow successful prediction. Rule selection is determined from an assessment of the spectral distributions.

4. K-nearest neighbor (KNN)

   The k-nearest neighbor (k-NN) algorithm classifies the objects by selecting the class of the closest training samples in the feature vector space. Amongst the other classifiers in the machine learning algorithms, k-NN is the simplest method. In the k-NN algorithm, an object is classified by a majority vote of its neighbors, and then the object is assigned to that class most common amongst its k nearest neighbors.

5. Support vector machines (SVM)

   Support vector machine is amongst the supervised learning methods that analyze

data and recognize patterns. It is generally used for classification and regression. Its concept based upon decision planes that define decision boundaries. A decision plane separates between a set of objects having different class memberships (Vapnic et al. 1995) and it generally produces better result than other well-known classifiers. There have been a few reasons why it outperforms many other classification methods:

i.      SVM finds a separating hyperplane in a way that maximizes the distance between it and the data points. So that the risk of misclassification is significantly reduced for an unseen data, compared to an arbitrarily selected separating hyperplane.

ii.     Data may not always be linearly separable within the given input space (almost-linearly-separable can be handled by adding slack variables to the constraints in the optimization problem). Transforming the data into a different space may help it to become linearly separable. The clever optimization method proposed by SVM uses kernels to make it possible to map the data to a very high dimensional space without actually paying too much cost, both in terms of memory space. To achieve this, instead of taking the data to a high dimensional space (e.g. a 30th order multidimensional polynomial of current space) and then computing inner products; it uses kernels to efficiently compute the inner products at that high dimensional space without actually transforming the data to that space. This even enables the SVM to compute inner products at infinite dimensional spaces.

iii.    SVM also does not consider the points that are far away from the separating hyperplane. The ones that are closer to hyperplane which are actually used in classification computation are called 'Support Vector's. By just using the closer ones, SVM increases computational efficiency while preserving classification perfomance, since far points do not give any information about the classification at hand, relatively to those that are close.

## 3.3 EXPERIMENTAL RESULTS

We tested the performance of the above features for the problem of view-independent face detection. The training set that is used consists of two sections for positive samples and negative samples. The positive samples in the training set are constructed images contain only faces. The face images in this class have been selected as the CVL face image dataset. On the other hand, the negative samples are constructed from two different sets of images. First set of images do not contain any skin regions but contain pixels that have color close to the skin color; second set of images contain skin regions that are not faces but other parts of the human body such as legs, arms, hands etc. Example non-face blobs are shown in Figure 3.9 and Figure 3.10. For the first part of the negative class, Compaq skin color dataset is used; however, for the second part, images have been collected from various sources like web pages and personal photos.

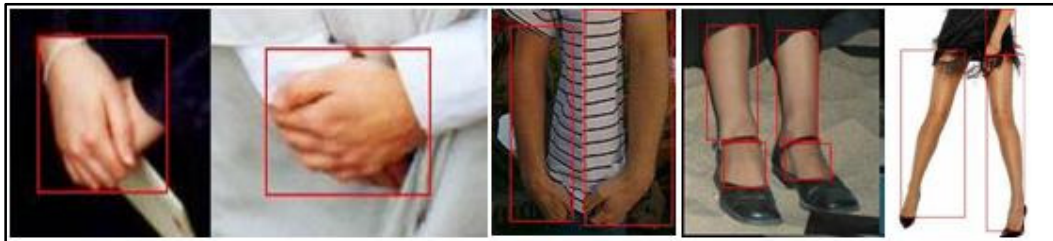**Figure 3.9: Images used for the first part of the negative class.**

**Images do not contain any skin regions**



Source: Compaq skin dataset

**Figure 3.10: Images that are used for second part of the negative class. Images**

**contain skin regions but no faces**

In order to observe the distributions of the extracted features in the training set, we sketched the feature vectors' histograms in Appendix 2. For the sake of clarity, all feature vectors are normalized to the [0-1] range. The moment histograms show that the distributions of the negative and positive features are quite different to discriminate feature space into two classes due to the positive feature histograms and their corresponding negative feature histograms' spread.

The accuracies of the facial features and classifiers are evaluated using several classifiers over the CVL dataset using leave one subject out (LOSO) cross validation. In LOSO, each subject is excluded from the training set in turn to be used as a test subject per iteration until all subjects are tested. Experimental results show that, using geometric and moment invariant features in a face detection task gives significantly accurate results. A summary of the results are shown in Table 3.6-3.9. In these tables, we compare face detection performances of methods using five different classifier implementations using PRTools (2011) and four different skin color detection methods. The acronyms used in the tables for classifiers: LB, QB, KNN-3, DT and SVM denote Linear Bayes Classifier, Quadratic Bayes Classifier, nearest neighbor classifier with 3 neighborhoods, Decision Tree, Support Vector Machine Classifier respectively, and for skin color detectors: M1, M2, SPM and SPM2 denote explicit method 1, explicit method 2, Skin Probability Map with 0.1 theta, Skin Probability Map with 0.2 theta respectively.

In Table 3.6, face detection results of geometric moment features on the CVL database are given. The highest detection rate of this feature set is obtained by using SPM and QB, which is 90.3%. Whereas the detection rate of the Viola Jones face detector on the same database (Table 3.10) is 43.3%, which is less than the given feature set. On the other hand Table 3.7 shows the detection rate of the R moment set. As we use R moments to learn face and non-face classes, detection rate increases to 93.3%, which is also better than the Viola Jones detection result. In Table 3.8, the results show the Zernike moment's detection rate. The maximum recognition rate of Zernike moments is 95.9%, which the best detection rate amongst the evaluated moment features. Lastly, in Table 3.9 gives the Flusser moments' results. The maximum detection rate of Flusser moments is 84.3%, which is the lowest detection rate amongst the moment set, though it

outperforms the VJ face detector on the CVL database. According to the detection rates given in Table 3.6 - Table 3.9 maximum detection rate is obtained from the Zernike moment features using SPM skin color detector and K-NN classifier.

In order to induce the detection rate, we also combined geometric shape features with moment features and experimental results are conducted on the best results of the moment features. As we can see from Table 3.11, maximum detection rate is obtained from the combination of the Zernike and geometric shape features, which is 95.98%. These combined features have introduced some improvement to the pure moment features.

**Table 3.6: Face detection results using geometric moment features**

| Geometric Moment Features | | | | |
|---|---|---|---|---|
| | M1 | M2 | SPM | SPM2 |
| | Detection Rate (%) | | | |
| LB | 71.8 | 69.7 | 72.5 | 71.8 |
| QB | 89.4 | 86.0 | 90.3 | 87.2 |
| KNN-3 | 86.4 | 81.4 | 86.8 | 84.0 |
| DT | 73.9 | 70.6 | 76.9 | 74.0 |
| SVM | 81.1 | 74.7 | 84.5 | 84.1 |

**Table 3.7: Face detection results using R moment features**

| R Moment Features | | | | |
|---|---|---|---|---|
| | M1 | M2 | SPM | SPM2 |
| | Detection Rate (%) | | | |
| LB | 76.6 | 71.8 | 79.0 | 79.1 |
| QB | 92.8 | 93.4 | 93.2 | 88.9 |
| KNN-3 | 85.6 | 78.9 | 84.8 | 83.9 |
| DT | 65.2 | 62.8 | 67.3 | 66.4 |
| SVM | 83.1 | 78.5 | 85.8 | 83.5 |

**Table 3.8: Face detection results using Zernike moment features**

| Zernike Moment Features | | | | |
|---|---|---|---|---|
| | M1 | M2 | SPM | SPM2 |
| | Detection Rate (%) | | | |
| LB | 88.3 | 83.4 | 89.9 | 85.5 |
| QB | 92.3 | 88.0 | 91.7 | 88.2 |
| KNN-3 | 95.6 | 91.0 | 95.9 | 92.0 |
| DT | 76.0 | 72.8 | 79.7 | 70.7 |
| SVM | 90.9 | 85.5 | 91.5 | 85.9 |

**Table 3.9: Face detection results using Flusser moment features**

| Flusser Moment Features | | | | |
|---|---|---|---|---|
| | M1 | M2 | SPM | SPM2 |
| | Detection Rate (%) | | | |
| LB | 65.4 | 64.8 | 67.6 | 60.9 |
| QB | 62.9 | 63.2 | 67.3 | 62.4 |
| KNN-3 | 80.4 | 73.1 | 83.9 | 77.6 |
| DT | 68.1 | 64.6 | 72.6 | 68.2 |
| SVM | 78.9 | 69.8 | 84.3 | 77.4 |

**Table 3.10: Results of the Viola Jones face detector on the CVL database**

| | True Positive | False Positive | Rate (%) |
|---|---|---|---|
| Viola Jones | 343 | 454 | 43.03 |

**Table 3.11: Face detection results using combined**
**features**

|  | LB | KNN-3 | DT |
|---|---|---|---|
|  | Detection Rates (%) | | |
| Shape + Geometric Moments | 91.77 | 94.98 | 86.82 |
| Shape + R Moments | 92.97 | 94.98 | 83.93 |
| Shape + Zernike Moments | 89.96 | 95.98 | 85.69 |
| Shape + Flusser Moments | 91.40 | 92.59 | 86.44 |

### 3.3.1 Conclusion

In this section, a method for view independent face detection is presented. The approach is based on segmentation of images into skin color blobs, which is followed by analyzing these skin blobs to extract some robust features regardless of their orientation, translation and scale. Finally these extracted robust features are used to train several well-known classifiers to differentiate faces from non-faces. Moment invariants, especially Zernike moments, has been quite successful for the face detection task. The utilization of moment invariants eliminates the need to scan the whole image at various scales to classify the regions whether they are face or non-face, which results in a computationally efficient face detector.

As it can be seen from the Table 3.6 to Table 3.11, face detection using shape features and moment invariants outperformed the well-known Viola-Jones face detector results on the CVL-Database, which contains both frontal and profile images. Although the Viola-Jones face detector works really well under conditions such as illumination variations and scale variations etc., it is not that accurate to detect non-frontal faces in

images. Our proposed algorithm, however, detects faces regardless of its orientation with high accuracy. Zernike moments with a K-NN classifier give the highest recognition rate (96%), which is not surprising since having orthogonal basis reduces the redundancy.

When moment features are combined with the geometric shape features the face recognition accuracy using LOSO cross-validation was almost the same as using Zernike moments only Table 3.11.

The proposed algorithm is dependent on the accuracy of the skin color classification; therefore any improvement of the skin is expected to improve the face detection performance.

# 4. FACIAL EXPRESSION RECOGNITION

Facial expression recognition is a challenging problem in computer vision and it has attracted a lot attention during the last decades due to the its potential application areas in human-computer interaction (HCI), forensics, health and e-learning. In human to human interactions emotions are expressed through any bodily motion but commonly through facial expressions. Although recognizing emotions via facial expressions is not only a trivial task, it is a key element in human communication since human faces can express emotions sooner than people speak or even realize their emotional states.

Facial expressions are generated by muscles that result in deformations around the eyes, nose, lips etc. These temporary deformations last for a few seconds. It is possible to recognize human emotions from visual data by analyzing these deformations. However, the problem is difficult due to the variations in appearance, lighting conditions and environmental settings. Most studies in literature present their results on posed databases. Recently, there is a growing interest in collecting spontaneous data in realistic settings as well.

In section 4.1, a brief literature survey on facial expression recognition and sparse representation-based classification is provided. In section 4.2, a facial expression recognition approach based on moment invariants and sparse representation-based classification is proposed. In section 4.3, we present a method for utilizing residuals of sparse representation-based face recognition for facial expression recognition.

## 4.1 LITERATURE SURVEY

### 4.1.1 Facial Expression Recognition

Various methods have been proposed during the last decades for automatic facial expression recognition (FER). These methods can be grouped into three main categories; methods that use appearance based features such as Gabor features, DCT features, and wavelets, model based methods such as labeled graphs, point distribution

model, optical flow etc., and hybrid methods that combine appearance and model based methods such as active appearance model (AAM).

Paul Ekman and his colleagues have performed extensive studies regarding human facial expressions (Ekman et al. 1970) and they found evidence to support universality in facial expressions. These *universal facial expressions* are those representing happiness, sadness, anger, fear, suprise and disgust. These prototypic emotional displays are also referred to as so called *basic emotions* that seem to be universal across human etnicities and cultures. Ekman and Friesen then develop the Facial Action Coding System (FACS) to code facial expressions where movements on the face are described by a set of action units (AU). The inputs of the system are still images of facial expressions often the peak of the expression. Ekman's work inspired many researchers to analyse facial expressions.

Mase (1991) used optical flow to recognize facial expressions. Lantinis et al. (1995) used a flexible shape and appearance model for facial expression recognition.

Currently, most of the existing facial expression recognition systems are based on 2D spatiotemporal facial features. The extracted features are either geometric features such as the shapes of the mouth, nose, eyes etc. and the location of facial salient points like corners of the eyes, mouth etc. or appearance features representing the facial texture including wrinkles, bulges and furrows. Some of the examples of geometric feature based methods are Chang et al. (2006) used a shape model defined by 58 facial landmarks, Pantic and Bartlett (2007) use a set of facial characteristic point around the mouth, eyes or nose, Kotsia and Pitas (2007) use a candidate grid and Erdem and Ulukaya (2011) used distance and angle based features using facial landmarks. On the other hand, an example of hybrid geometric and appearance features based methods is proposed by Tian et al. (2005). They used facial components' shapes and the transient features like crow-feet wrinkles and nasal-labial furrows.

There are a few approaches for facial expression analysis based on 3D face models such as the method proposed by Huang and colleagues (2004; 2003) who used features extracted by a 3D face tracker that is called as "Piecewise Bezier Volume Deformation

Tracker" (Cohn et al. 2003; Tao and Huang 1999) and focused on the analysis of head movement based upon a cylindrical head model.

## 4.1.2 Sparse Representation-Based Classifier

Sparse representations of digital signals have become very popular in recent years and utilization of this representation in a classification problem means searching for the most convenient representation of a signal in terms of linear combination of features in an overcomplete dictionary shows that it can be exploited for the problem of classification. Compared to other classification methods, sparse representation generally offer better performance under the condition of an overcomplete dictionary.

Using sparse representation in a classification problem can be considered as a generalization of nearest neighbor (NN) or nearest subspace (NS). The key point in the NN classifier is classification based on the best representation of the test vector according to a single training vector. The NS classifier, however, classifies the test vector using the best linear representation principle according to the whole training vectors in per class. The usage of sparse representation in the classification problem on the other hand is to represent the test object as a linear combination of the dictionary from the same object class while keeping the signal reconstruction error as less as possible. General description of classification based on sparse representation can be found in the following section.

The classification problem of finding sparse representation of a test sample in a given dictionary can be formulated as follows. Given an $N \times M$ matrix A, which contains the given $n_i$ training samples of the i-th class as in columns like $A_i = [v_{i,1}, v_{i,2}, ..., v_{i,n_i}]$, the matrix A is generally $M \gg N$ and the test sample is $y \in R^N$. The problem of sparse representation is to find a vector $x$, such that $y = Ax$ and $\|x\|_1$ is minimized;

$$(\ell^1): \quad \hat{x}_1 = argmin\|x\|_1 \quad subject \ to \ Ax = y \tag{4.1}$$

where $\|x\|_1$ is the $\ell^1$ norm. After calculation of the optimal vector $\hat{x}_1$ in $M \times 1$ dimension in terms of (4.1), it should be used to classify the test vector y into one of the

K object classes. Let $\delta_k(\hat{x}_1)$ be a new vector whose only non-zero elements are the elements in $\hat{x}_1$ that are associated with class k. Using only these non-zero coefficients associated with the k-th object class, it can be approximated the given test sample y as $A\delta_i(\hat{x}_1)$. Then the test sample y can be classified to class $i$ using this approximation to the object class that minimized the residuals between y and $A\delta_k(\hat{x}_1)$.

$$argmin_i \, r_i(y) = \|y - A\delta_i(\hat{x}_1)\|_2 \; for \; i = 1,2,..,k \qquad\qquad \textbf{(4.2)}$$

The aforementioned algorithm is summarized below;

**Algorithm 4.1: Sparse representation based classification**

---

1: **Input:** a matrix of training samples A=$[A_1, A_2, .. .. .. , A_k] \in$ $\mathbb{R}^{mxn}$ for k classes, a test sample $y \in \mathbb{R}^m$,

2: Solve the $\ell^1$-minimization problem:

$$(\ell^1): \quad \hat{x}_1 = argmin\|x\|_1 \; subject \; to \; Ax = y$$

3: Compute the residuals $r_i(y) = \|y - A\delta_i(\hat{x}_1)\|_2$ for i = 1,2,..,..,..,k

4: **Output:** identity(y) = $argmin_i \, r_i(y)$

---

*Source*: Wright et al. (2009)

Wright et al. (2009) adapted this new classification technique to the face recognition problem and experimental results showed that if the sparsity is properly harnessed, recognition rates are significantly high. This motivated the use this classification technique for the facial expression recognition problem.

## 4.2 FACIAL EXPRESSION RECOGNITION USING MOMENT INVARIANTS AND SPARSE REPRESENTATION-BASED CLASSIFIER

In a facial expression recognition system, it is very important to choose an appropriate feature space that represents the emotional expressions as much as possible. The performance of the chosen feature space directly affects the classification accuracy. Recently, moment based feature extraction methods have been widely used in the object recognition task due to their ability to capture both global and detailed geometric information about an image or an image region. We used them for the face detection task, as was presented in the previous chapter and we observed that these statistical moments, especially Zernike moments, represent the facial features quite well.

There are a few studies that utilize Zernike moments to extract emotional features for a facial expression recognition system. Jiang and Shu (2004) used Zernike and geometric moments to recognize happiness and sadness, Fang et al. (2010) used Zernike moments to recognize infant facial expressions and Lajevardi and Hussain (2010) used Zernike moments and a naive Bayes classifier. Experimental results of the mentioned researchers are promising when high order Zernike moments are used.

In this section, we show that using Zernike moments as features for rotated expressive face images outperform using the original pixel intensity features when a sparse representation-based classifier is used.

### 4.2.1 Feature Extraction For Facial Expression Recognition

Unlike the other descriptors such as Fourier, DCT, Gabor features, Zernike moments do not require closed boundaries to capture information from images. However, due to the fact that any unrelated pixels in the background affect the recognition ratio, we extract face locations from the images. In order to estimate face locations, we used landmark points which come together with the database. These landmark points are shown on an example emotional image in Figure 4.1.

As it can be seen from Figure 4.1, the face is surrounded by 68 landmark points, which we use to determine the face region. In order to localize the face; first, maximum and

minimum x coordinates were found to determine the left and right boundaries of the face. According to the Figure 4.1, these maximum and minimum points are the 1$^{st}$ and 16$^{th}$ landmark points. A similar approach is followed to determine the upper and lower boundaries of the face. We did not align the images since Zernike moments are rotation and scale invariant. A typical result of an extracted face images is shown in Figure 4.2.

**Figure 4.1: An example image with landmark points**
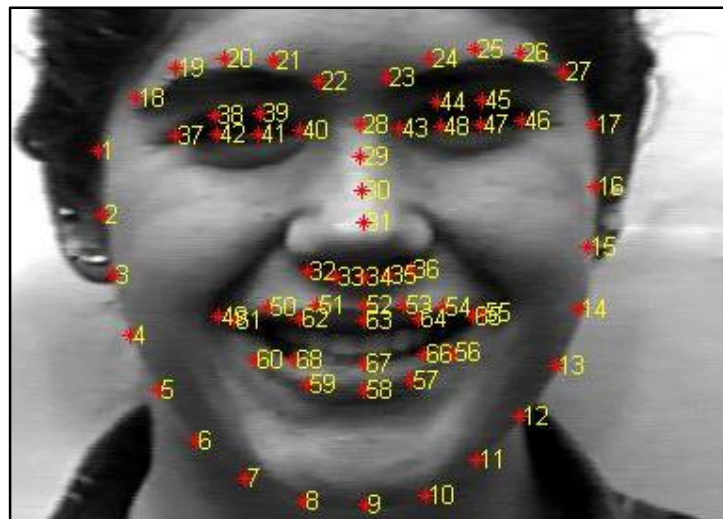


*Source*: CK+ database

**Figure 4.2: A typical extracted image**



*Source*: CK+ database

Due to the facial biometry, the height of a face is longer than its width. However; the kernel of Zernike moments is based on orthogonal polynomials defined in the polar coordinate space inside a unit circle (Lajevardi and Hussain 2010), therefore the images should have equal width and height. In order to satisfy this condition, we resized the images into squares using their longer edges. Typical extracted and resized images are shown in Figure 4.3. Resizing is a necessary process for Zernike moment extraction due to its basis.

We also used another technique to extract face region. In the second technique, many of the background pixels were eliminated, since every irrelevant pixel in the background adds noise to the feature space. In order to get rid of the background, we used the landmarks $1^{st}$ to $17^{th}$. These seventeen points helped us to eliminate the pixels under the chin and cheek. On the other hand; we drew a half circle centered at $28^{th}$ point. The diameter of the circle was calculated from the distance between the $1^{st}$ and $17^{th}$ landmark points. The half circle mostly eliminates the redundant pixels which do not carry any information about the emotional expression. An example image can be seen in Figure 4.4.

**Figure 4.3: Resized (left) and extracted (right) images**



*Source*: CK+ database

**Figure 4.4: Face region without background**

After the images were preprocessed, we extracted two types of features, which are Zernike moments and original pixel intensity values. However, the images are quite big to use original pixel intensities as features; therefore, we subsampled the images to size 64x64, and then these pixels in the image were used as 4096-D features.

### 4.2.2 Experimental Results

Experiments were done using the Extended Cohn-Kanade (CK+) database. For person independence, we used leave one subject out (LOSO) cross-validation. Both Zernike moments and original pixel intensity features are used as features to represent the emotional state of a test face image, and these features are classified using a sparse representation based classifier. In order to use sparse representation based classification, we organized both Zernike and original pixel intensity features as columns of matrices A and B respectively, where each column vector represents a training image has with dimension $m \times 1$ (36 for Zernike, 4096 for original pixel intensity), then we found the best linear representation of the test vector with respect to the dictionaries A and B using the Algorithm 4.1.

As we described in the previous chapter, Zernike moments are invariant to rotation. Original pixel intensities, however, are not. Even if the images are rotated with an arbitrary angle, the magnitudes of the Zernike moments remain the same; therefore, the accuracy of the facial emotion recognizer is not affected much. On the other hand utilization of original pixel intensities as features resulted in a low recognition rate for rotated images. Furthermore, experiments showed that although the accuracy of the original pixel intensities is higher than the Zernike moment features in the absence of rotation, Zernike moments outperformed the original pixel intensity features under the presence of rotation with an arbitrary angle between -45 to +45 degrees.

We tested our algorithm on the images with one of the six basic emotions anger, disgust, fear, happiness, sadness and surprise. In order to reduce the effect of illumination variations, we used histogram equalization. Below, the experimental results are given in Table 4.1.

Experimental results given in Table 4.1 shows that pixel based features gives better facial expression recognition rates in the absence of rotation, which is 83%, on the other hand with Zernike moment features it is 74%. However; as can be seen from theTable 4.1, in the presence of rotation, the recognition rate with original pixel intensity features falls to 69%, whereas, the recognition rate of Zernike moment features remains unchanged (74%), meanings Zernike moment features outperform the original pixel intensity features. Notice that, both feature extraction methods give its best performances with the exact face region method and with histogram equalization. Besides, we can see from the Table 4.1 that the overall accuracy of the exact face images are better than the face rectangle results. This is not surprising since eliminating the background pixels as much as possible eliminates the noise from the feature space regardless of feature type.

**Table 4.1: Expression recognition results**

| Face Extraction Method | Feature Extraction Method | |
|---|---|---|
| | Pixel Intensity Features | Zernike Moment Features |
| Exact Face Region | 81% | 65% |
| Exact Face Region + Histogram Equalization | 83% | 74% |
| Exact Face Region + Arbitrary Rotation Angle | **58%** | **65%** |
| Exact Face Region + Histogram Equalization + Arbitrary Rotation Angle | **69%** | **74%** |
| Face Rectangle | 78% | 69% |
| Face Rectangle + Histogram Equalization | 79% | 71% |
| Face Rectangle + Arbitrary Rotation Angle | **59%** | **70%** |
| Face Rectangle + Histogram Equalization + Arbitrary Rotation Angle | **59%** | **71%** |

### 4.2.2.1 Conclusion

In this section of thesis, we evaluated two kinds of feature extraction methods for rotation invariant facial emotion recognition. First one is the original pixel intensity values, and second one is Zernike moments.

The experimental results conducted on the CK+ database showed that, when original pixel values are used in a facial expression recognition system the facial expression recognition rate is around 83%. However; when pixel values are used as features, any

geometric variations of the image affect the recognition ratio significantly, and result in a low recognition rate 69%.

The second feature type called Zernike moments, as a course of its nature, are invariant to geometric variations. In our experiments, although it did not produce the best result, it gave better results than pixel intensity features under the presence of rotation. Considering the realistic scenarios, humans benefit from the head gestures to express their emotions. The head gestures, however, cause the faces to rotate with an arbitrary angle. Zernike moments are able to tackle this problem. The recognition rates show that maximum recognition rate of the Zernike moments is 74%, and when rotation is applied to images, result does not change, which show that Zernike moments can produce reliable features regardless of the orientation of the images.

As a result, the performance of Zernike moments in a FER system surpasses the performance of original pixel intensity features under the presence of rotation. The superiority of the Zernike moments to the pixel intensity features is their robustness of the geometric variations.

## 4.3   UTILIZING RESIDUALS OF SPARSE REPRESENTATION-BASED FACE RECOGNITION FOR FACIAL EXPRESSION RECOGNITION

The idea of utilizing sparse representation based facial expression recognition has intrigued researchers recently due to the utilization of efficient convex optimization algorithms, when the linear representation is sparse enough. Discriminating emotions from images using sparse representation is motivated by the accuracy of such methods in face recognition problem (Wright et al. 2009). Some researchers exploited the Sparse Representation based Classifiers (SRC) in a FER system; Zafeiriou and Petrou (2010) used difference images (calculated by subtracting the neutral image intensities from the corresponding intensities of fully expressive facial expressive image called apex image), and also they studied facial image grids in order to discriminate facial emotions; Wang et al. (2010) compared the sparse representation with various traditional algorithms such as 2D PCA and curvelet transform; Cotter (2010) showed that the performance of SRC is superior to the Gabor based features when the test images are occluded.

In this section, we show that the sparse representation of expressive facial images from a basis set of neutral face images can be utilized in a facial expression recognition system. Intuitively, we show that the *non-representable* part of the face images called residual images can be a good approximation to the difference images. This technique could be used when the true neutral face is not known as difference image estimation. Experimental results conducted on CK+ database are promising without any complicated process of facial feature extraction and tracking.

### 4.3.1 Outline Of The Proposed Algorithm

In this section, we proposed a combined face and facial expression recognition method using two-stage sparse representation based classifier (SRC). Zafeiriou and Petrou (2010) was proposed the difference of the expressive face and its corresponding neutral face as a feature set for the FER system since they stated that the expressive face of a person contains strongly identity related cues. Therefore they suggested that the difference image approach to eliminate the facial identity in a certain degree. However, this approach requires that the identity of the person is known and the neutral face of that person is available.

In this part of the thesis, we propose a two-stage sparse representation based method to estimate the difference image using an initial face recognition step, when given only an expressive face image. First, we perform face recognition using the neutral faces of all subjects as the dictionary. At this stage, the expressive face is represented as a linear combination of neutral faces in the overcomplete dictionary of neutral faces. Then, the difference image is estimated using the residual image, which is the result of the face recognition stage. The residual image is then used for the test image of the facial expression recognition step.

### 4.3.2 Utilizing Residuals Images For Facial Expression Recognition

In order to apply the sparse representation-based classification algorithm for facial expression recognition, the dictionary needs to include all facial expressions of the individuals and the given test image is typically represented with an image of the same

person with the same expression; that is, a smiling face of an individual can be best represented by another face of the same individual and the representation would improve even better with another smiling face of that individual. The fundamental issue with this approach is that sparse representation of the emotional image tents to give high responses for images of the same person that depicts a different facial expression (Zafeiriou and Petrou 2010). Therefore, facial expression recognition cannot be accomplished when the dictionary does not have a smiling face of that individual. Hence, identity related cues should be removed from the image, in order to resemble to another person's smiling face.

### 4.3.2.1  Stage 1: Face recognition and estimation of the residual images

We assumed that an emotional face image is available and we will classify this image into one of the six facial expressions. However, as a course of the emotional images carry both identity and emotional related information. Thus, we have to eliminate the identity related information as much as possible. One way to eliminate identity of an emotional image is to subtract the neutral image of the same person from the expressive image. However, the neutral face image may not know always be available, since we may not know the identity of that person. In order to estimate the neutral face images, we have to find the sparse representation of a given expressive image with regards to the dictionary A, which consists of the neutral images belong to the all subjects. Intuitively, this stage corresponds to a face recognition system. Then the residual images are obtained during this stage, which contain minimum amount of identity while describing of emotional knowledge showed in Figure 4.5

**Figure 4.5: Stage 1: Estimating residual image**
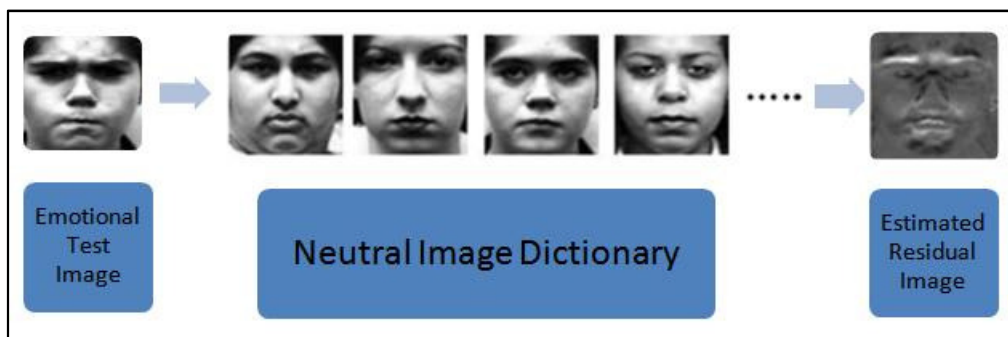
In stage 1, we define the overcomplete dictionary as $A = [A_1, A_2, ..., ..., A_k] \in \mathbb{R}^{mxn}$, where $A_i$ is matrix containing at least one neutral face image of person $i$ in its columns. Then, sparse representation is obtained by solving Algorithm 4.1. However, to obtain the residual image of the related person that will be used in the second stage as an input; we normalized the coefficient of the winning class decomposition coefficients. The normalization is a necessary process since Algorithm 4.1 finds the representation of the expressive image with respect to the whole neutral image dictionary. However; only the coefficients in the representation that are associated by the winning class are considered to estimate residual image of a person; hence, this results in lots of information loss.

### 4.3.2.2  Stage 2: Facial expression recognition using residual images

Given the residual images from stage 1 as input, we want to find the sparse representation of the residual image with regards to the dictionary B, which consist of full set of difference images except one being tested. We defined the overcomplete dictionary as $B = [B_1, B_2, ..., ..., B_6] \in \mathbb{R}^{mxk}$, where $B_j$ is a matrix containing the difference images of emotion $j = 1,2,3, ..., 6$ for all training subjects (except one being tested). Then the sparse representation is obtained by solving Algorithm 4.1, the recognized expression is: emotion $(z) = argmin_j r_j(z)$. The algorithm of the stage 2 is depicted in Figure 4.6.

**Figure 4.6: Stage 2: Recognizing emotion**
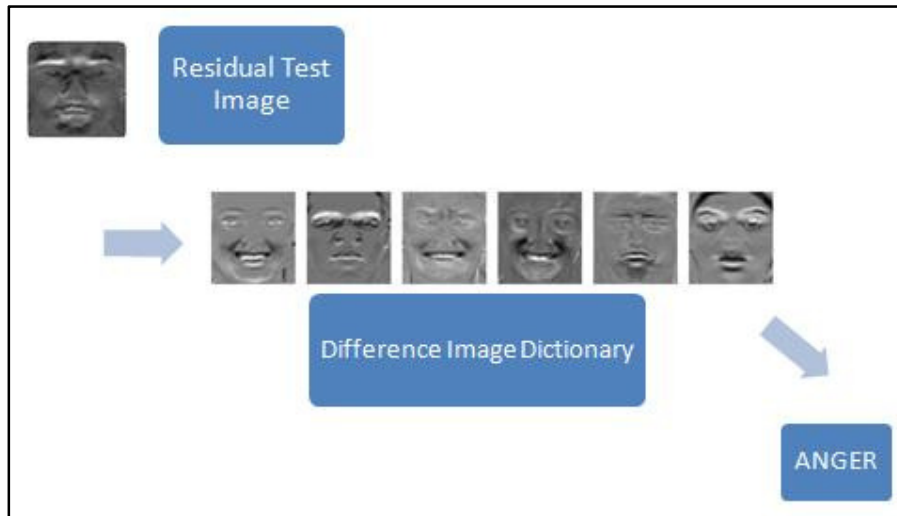
The steps of the proposed facial expression recognition algorithm are summarized in Algorithm 4.2.

**Algorithm 4.2: Proposed expression recognition algorithm**

1: **Inputs:** A training matrix $A = [A_1, A_2, \dots, \dots, A_k] \in \mathbb{R}^{m \times n}$ containing neutral face images in columns for k subjects, a test sample $y \in \mathbb{R}^m$, which is an expressive face of one of the k subjects.

2: Solve the $\ell^1$-minimization problem:
$$\hat{x}_1 = argmin\|x\|_1 \quad subject\ to\ Ax = y$$

3: Compute the residuals $r_i(y) = \|y - A\delta_i(\hat{x}_1)\|_2$ for i = 1,2,..,..,..,k.

4: Identity $(y) = i^* = arg\ min_i\ r_i(y)$.

5: Find the normalized weight vector $\lambda(\hat{x}) = \dfrac{\delta \cdot (\hat{x})}{\sum_j (\delta \cdot (\hat{x}))_j}$ .

6: Compute the residual image z = y - A$\lambda(\hat{x})$

7: Form the training matrix for facial expression recognition $B = [B_1, B_2, \dots, \dots, B_6] \in \mathbb{R}^{m \times k}$, containing the difference images of the six basic emotions for all training subjects (except the subject being tested for subject independence, which is $i^*$).

8: Solve the $\ell^1$-minimization problem:
$$\hat{w}_1 = argmin\|w\|_1 \quad subject\ to\ Bw = z$$

9: Compute the residuals $r_j(z) = \|z - B\delta_i(\hat{w}_1)\|_2$ for i = 1,2,..,..,..,6

10: **Output:** The recognized emotion is: emotion$(y) = j^* = arg\ min_j r_j(w)$

## 4.4 EXPERIMENTAL RESULTS

Experimental results are conducted on the CK+ database, which consists of sequences in which the first frame of the sequences are the neutral images, and last frame of the sequences are expressive images. Expressive images are separated into 6 different emotional classes called anger, disgust, fear, happiness, sadness and surprise. Some of the expressive images can be seen in Figure 4.7.
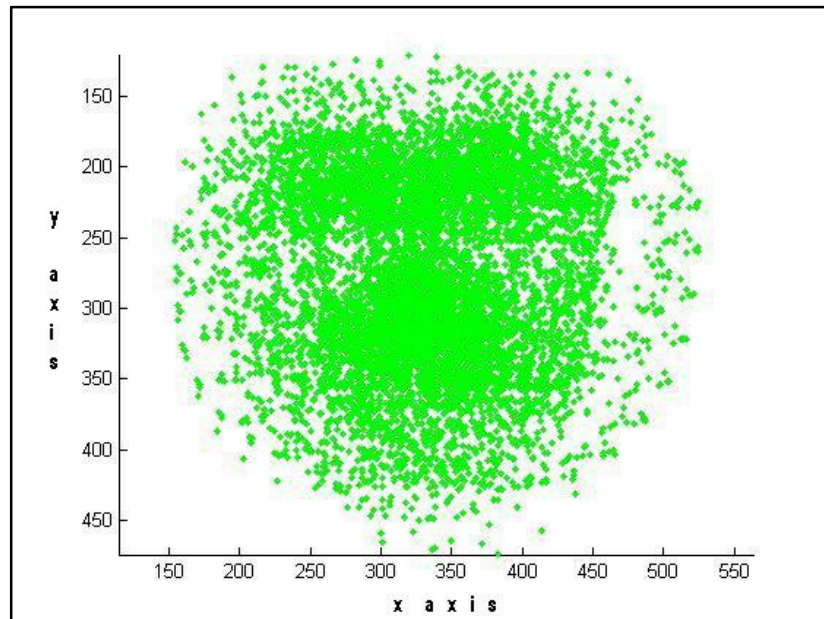
**Figure 4.7: Expressive images**



*Source*: CK+ database

In order to use the CK+ database; we aligned, cropped and resized the images since the landmark points have different orientation according to the axis, Figure 4.8.

**Figure 4.8: Landmark points without alignment**

Alignment of the images is performed using the landmark points. First, we rotated the images using the angle between the inner points of the eyes in order to compensate in-plane head orientation; second, we scaled the images so that the sum of distances of the three point pairs such as $40^{th}$-$43^{rd}$, $1^{st}$-$17^{th}$ and $2^{nd}$-$16^{th}$ is constant; then we cropped the images so that the nose tip is at the center of the images.

The landmark points after the alignment step can be seen in Figure 4.9. Finally, we sub-sampled the images to a size of 64×64 to reduce the feature dimension to 4096-D.

**Figure 4.9: Landmark points after alignment**
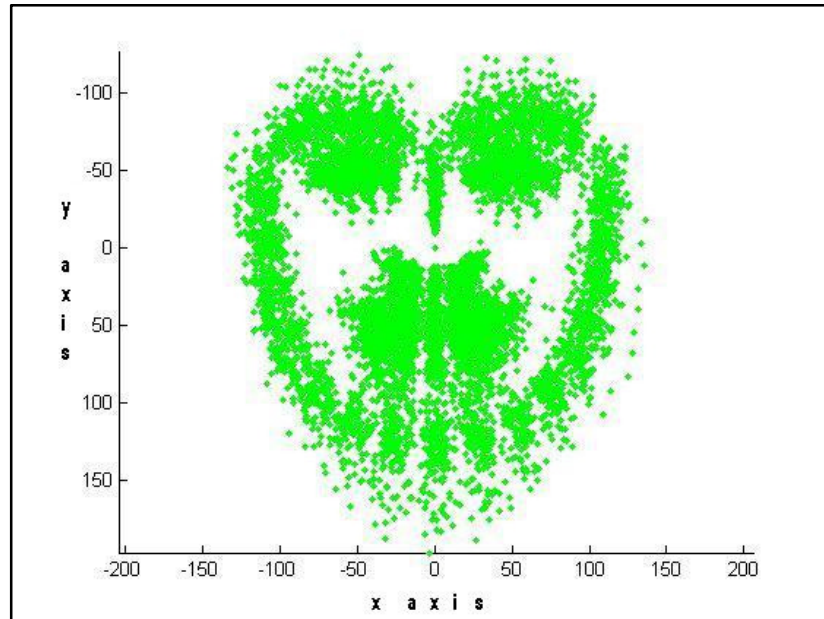


In the first stage, in order to find the residual images as difference image estimators, we used neutral images as A, which is the dictionary of face recognition part. The results of this stage of the proposed algorithm are illustrated in Figure 4.10. The left images show the input expressive images, the middle images show the neutral images with the maximum coefficient from the recognized class, and the right images show the estimated residual images (a anger, b disgust, c fear, d happiness, e sadness, f surprise). We can clearly observe that residual images contain mostly facial expression related information. A sample decomposition vector is given in Figure 4.11, which shows that

the maximum coefficient indicates the correct class of the person. The overall accuracy of the stage 1 (face recognition stage) is 88%.

The residual images that are yielded by stage 1 are given as input to stage 2, which is the emotion recognition part. A sample decomposition vector shown in Figure 4.12

**Figure 4.10: Results of the first stage**
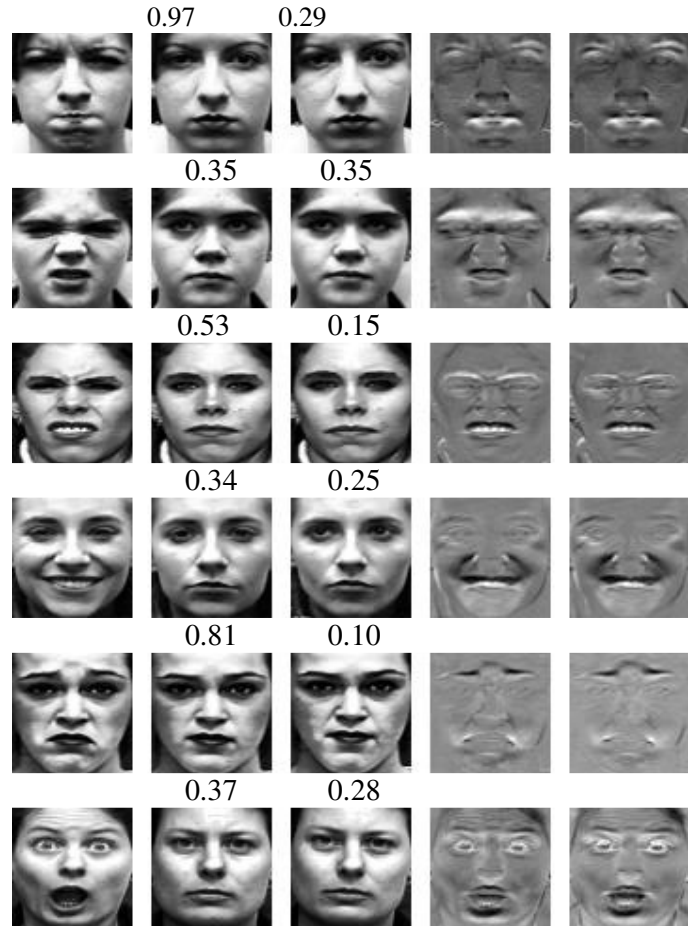


In Fig. 4.10, the images in the first column are the expressive images that those are used as inputs to the system. The images in the second and third column are the neutral images of the winning class with max two coefficients that are shown above the images. The forth column images are the estimated residual images and the last column images are the corresponding difference images.

**Figure 4.11: A sample decomposition vector for stage 1**
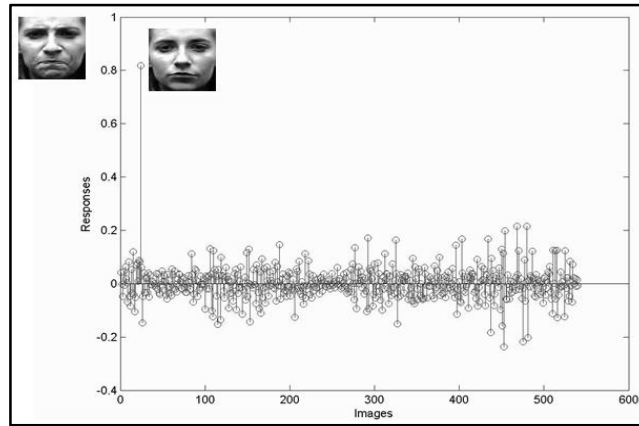


**Figure 4.12: A sample decomposition vector for stage 2**
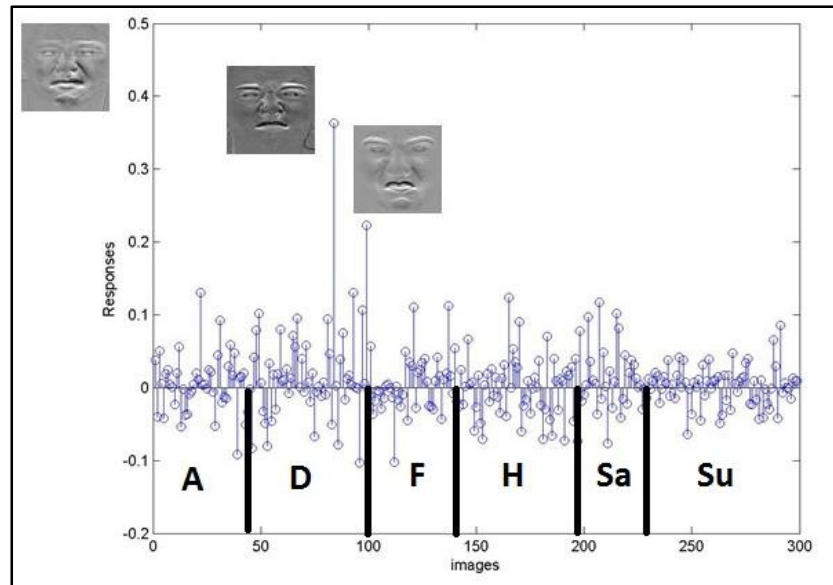


It can be observe from the Figure 4.12, many of the high responses fall into the same expression class, which is disgust. The delimiters show the training set labels, (A for anger, D for disgust, F for fear, H for happiness, Sa for sadness, Su for surprise). The confusion matrix for the emotion recognition stage is given in Table 4.2. The rows depict the actual expressions of the image and the columns depict the predicted

expressions of images. As we can see from the table, happiness, surprise and disgust expressions are recognized with a high accuracy but anger and fear's accuracies are relatively weak from the others. According to the Table 4.2, fear expression is mostly confused with happiness and surprise since these expressions include mostly mouth related changes. On the other hand, anger is confused mostly with disgust. However; the overall emotion recognition rate is 82%, which is quite promising. This accuracy falls to 78% when the difference images is estimated using the neutral image with the maximum coefficient in the winning class, which shows that the residual computed by the sparse representation works better than picking a single neutral image.

Table 4.2: The confusion matrix (in %) for the emotion recognition stage

|  | Anger | Disgust | Fear | Happiness | Sadness | Surprise |
|---|---|---|---|---|---|---|
| Anger | 58 | 20 | 0 | 7 | 2 | 13 |
| Disgust | 0 | 90 | 0 | 3 | 0 | 7 |
| Fear | 0 | 0 | 28 | 32 | 8 | 32 |
| Happiness | 0 | 0 | 0 | 98 | 0 | 1 |
| Sadness | 4 | 0 | 0 | 4 | 75 | 18 |
| Surprise | 0 | 0 | 0 | 4 | 0 | 96 |

### 4.4.1 Conclusion

In this section, we showed that the non-representable part of the expressive face images named residual images or the reconstruction error from the representation from neutral bases can be a good approximation to the difference images. In a FER system, a good estimator for difference images is useful in that they reveal mostly expression specific but not identity related information. When the neutral face image is not given for a given expressive test image, the proposed technique can be effectively used for estimating its difference images.

The experimental results conducted on CK+ database using leave one subject out approach in facial expression recognition part give an average emotion recognition rate of 82%, which is quite promising. The main advantage of the proposed algorithm is that at first stage aims directly at computing residual/difference image but not directly at

face recognition. As a matter of fact, when the correct identity is estimated by the same approach and then the difference image is computed by subtracting the best-matching (high peak image of the classified person) of that person from the test image, average recognition rate falls to 78%

# 5. CONCLUSION AND FUTURE WORK

Skin color is a useful cue for many applications due to its robustness to geometric variations. However, choosing a proper color space for segmentation and choosing a proper classification method are not easy problem. In this thesis, we evaluated and analyzed several skin color detection algorithms, which use the RGB and $YC_rC_b$ color spaces and four different classifiers. Experimental results conducted on Compaq skin dataset showed that Method 2 (thresholding on RGB color space) gave the best result with respect to the true positive rate. On the other hand, in terms of elimination of unwanted regions of the images like eyebrows, lips, the $YC_rC_b$ gave the best results since its false positive rate is the lowest.

Although face detection has been a widely studied problem, detection of non-frontal faces in images is still a challenging task. We proposed a new face detector that is able to detect non-frontal faces, which begins with segmentation of images into skin color blobs, and then follows by computation of scale, translation and rotation invariant features to learn a statistical model of faces and non-faces. The person independent experimental results conducted on the CVL database showed that the face detection rate reaches 95% while the well known Viola Jones face detector gives a detection rate of 43%.

Another important task for human-computer interaction is emotion recognition, which is a challenging problem. In this thesis, we proposed the utilization of Zernike moments to classify facial expressions into six basic emotion classes: anger, disgust, fear, happiness, sadness and surprise. In this thesis, we compared the performances of Zernike moments and original pixel intensities as expressional features under the presence of arbitrary rotations and we used a sparse representation based classifier to classify facial expressions. The experimental results on the CK+ database gave an average emotion recognition rate of 74% with Zernike moment invariants. Whereas, when we used the pixel intensities as features, the average emotion recognition rate fell to 69%.

We also showed that the non-representable part of the expressive face images named residual images or the reconstruction error from the representation from neutral bases

can be a good approximation to remove identity related information from the images. Then, we stated that when the neutral face image is not given for a given expressive test image, the residual images can be effectively used as difference images. Experimental results conducted on CK+ database using LOSO gave promising results, which is 82%. On the other hand, when the correct identity is estimated by the same approach and then the difference image is computed by subtracting the best-matching (high peak image of the classified person) of that person from the test image, average recognition rate falls to 78%. The fundamental advantage of the proposed algorithm is that when the true neutral face image is not give for a given test image, this technique can be used effectively to estimate its difference image.

Possible directions for future research may be listed as follows:

i.   Improving the accuracy of skin color segmentation methods since face detection rate directly depends on the accuracy of skin color segmentation.

ii.  Improving the computation time using graphical processing unit (GPU) since many applications require real-time operation.

iii. In order to extend the work in Section 4.4 we want to construct a large enough neutral face image dictionary, and compute the residual image, which may allow facial expression recognition of test images of individuals that have not seen before

# REFERENCES

***Books***

Darwin C., 1890. *The expression of the emotions in man and animals*, 2nd edition. London : John Murray.

Ekman P., 1982. *Emotion in the human face*, 2nd edition. New York : Cambridge University Press.

James W., 1890. *The principles of psychology*. New York : Henry Holt.

Ratsch G & Meir, R., 2003. *An introduction to boosting and leveraging*. S. Mendelson and A. J. Smola Ed., Advanced Lectures on Machine Learning, Berlin Heidelberg : Springer-Verlag, pp. 118-183.

Tian Y.L., Kanade T. & Cohn J.F., 2005. *Facial expression analysis*. Handbook of Face Recognition, S.Z. Li and A.K. Jain, eds., pp. 247-276, Springer.

Zitova B., Suk T. & Flusser J., 2009. Moment and moment invariants in pattern recognition. John Wiley.

*Periodicals*

Bishop C. and Viola, P., 2003, Learning and vision: discriminative methods. *In ICCV Course on Learning and Vision*.

Brand J., and Mason, J., 2000, A comparative assessment of three approaches to pixel level human skin-detection. *In Proc. of the International Conference on Pattern Recognition*, vol. 1, pp. 1056-1059.

Celebi M. E. and Aslandogan, Y. A., 2005. A comparative study of three moment-based shape descriptors, *Proc. Of the International Conference of Information Technology: Coding and Computing*, pp. 788-793.

Chang Y., Hu, C., Feris, R. and Turk, M., 2006. Manifold based analysis of facial expression. *J. Image and Vision Computing*, vol. 24, no. 6, pp. 605-614.

Chen Q., Wu, H., and Yachida, M., 1995. Face detection by fuzzy pattern matching. *In Proc. of the Fifth International Conference on Computer Vision*, pp. 591-597.

Cohen L., Sebe, N., Garg, A., Chen, L., and Huang, T., 2003. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding*, vol. 91, nos. 1-2, pp. 160-187.

Crow F., 1984. Summed-area tables for texture mapping. *In Proc. of SIGGRAPH,* vol. 18, pp. 207-212.

Ekman P., and Friesen, W. V., 1971. Constants accross cultures in the face and emotion. *Journal of Personality and Social Psychology*, **17**(2):124-129.

Erdem C.E., Ulukaya, S., 2010. Duygu tanımlama için geometric yüz özniteliklerinin karşılaştırılması. *IEEE 19th Signal Processing and Applications Conference*, (SIU) Antalya, Turkey.

Fang C.Y., Lin, H.W. and Chen, S.W., 2010. An infant facial expression recognition system based on moment feature extraction. *VISAPP 2010, International Conference on Computer Vision Theory and Applications*.

Fleck M., Forsyth, D. A., and Bregler, C., 1996. Finding naked people. *In Proc. of the ECCV*, vol. 2, 592-602.

Gang J., Sung E., 1999. Frontal-view face detection and facial feature extraction using morphological operations. *Pattern Recognition Letters*, pp. 1053-1068.

Haj M. A., Bagdanov, A. D., Gonzàlez, J., Roca, X. F., 2009. Robust and efficient multipose face detection using skin color segmentation. *IbPRIA '09 Proceedings of the 4th Iberian Conference on Pattern Recognition and Image Analysis*.

Hjelmas E. and Low. B. K., 2001. Face detection: A survey", *Computer Vision and Image Understanding*, 83: pp. 236-274.

Hsu R. L., Abdelmottaleb M., and Jain, A. K., 2002. Face detection in color images. *IEEE Trans. Pattern Analysis and Machine Intelligence* **24**(5), pp. 696-706.

Hu M.K, 1962. Visual pattern recognition by moment invariants. *IRE Trans. Information Theory*, vol. IT-8, pp. 179-187.

Huang T. S. and Yang G., 1994. Human face detection in complex background. *Pattern Recognition*, vol. 27, no. 1, pp. 53-63.

Jiang L. and Shu H., 2004. Comparison of hu moment and zernike moment in application of facial expression recognition. *Journal of Luoyang University*, vol. 19, no. 2, pp. 14-17.

Jones M. J., and Rehg J. M., 1999. Statistical color models with application to skin detection. *In Proc. of the CVPR '99*, vol. 1, pp. 274-280.

Jordao L., Perrone M., Costeira J. and Santos Victor J., 1999. Active face and feature tracking. *In Proceedings of the 10th International Conference on Image Analysis and Processing*, pp. 572-577.

Khotanzad A., Hong Y. H., 1990. Invariant image recognition by zernike moments. *IEEE Trans. Pattern Anal. Machine Intell.*, 12, pp. 489-498.

Kim Yong-Sung, Kim, Whoi-Yul, 1998. Content-based trademark retrieval system using visually salient feature. *J. Image Vision Comput*. 16, pp. 12-13.

Kotsia I. and Pitas I., 2007. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Trans. Image Processing*, vol. 16, no. 1, pp. 172-187.

Lajevardi S.M. and Hussain Z.M., 2010. Higher order orthogonal moments for invariant facial expression recognition. *Digital Signal Processing*, 20, pp. 1771-1779.

Lanitis A., Taylor C. and Cootes T., 1995. A unified approach to coding and interpreting face images. *In Proc. International Conf. on Computer Vision*, pp. 368-373.

Lee J. Y. and Yoo S. I., 2002. An elliptical boundary model for skin color detection. *In Proc. of the 2002 International Conference on Imaging Science, Systems, and Technology*.

Leung T.K., Burl M.C. and Perona P., 1995. Finding faces in cluttered scenes using random labeled graph matching. *Proc. Fifth IEEE Int'l Conf. Computer Vision*, pp. 637-644.

Liu J., Liu Y. and Yan C., 2008. Feature extraction technique based on the perceptive invariability. *Proceedings of the Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, Shandong, China, pp. 551-554.

Mase K., 1991. Recognition of facial expression from optical flow. *IEICE Trans.* **E74**(10), pp. 3474-3483.

Mckenna S., Gong S. and Raja Y., 1998. Modelling facial color and identity with gaussian mixtures. *Pattern Recognition,* **31**(12), pp. 1883-1892.

Menser B. and Wien M., 2000. Segmentation and tracking of facial regions in color image sequences. *In Proc. SPIE Visual Communications and Image Processing 2000*, pp. 731-740.

Ogihara A., Shintani A., Takamatsu S., Igawa S., 1996. Speech recognition based on the fusion of visual and auditory information using full-frame colour image. *IEICE Trans. Fundamentals*, pp. 1836-1840.

Pantic M. and Bartlett, M.S., 2007. Machine analysis of facial expressions, Face Recognition, K. Delac and M. Grgic, eds., pp. 377-416, *I-Tech Education and Publishing*.

Peer P., Kovac J. and Solina, F., 2003. Human skin colour clustering for face detectio. *In submitted to EUROCON 2003 – International Conference on Computer as a Tool*.

Ryan A., Cohn J., Lucey S., Saragih J., Lucey P., la Torre, F. D. and Rossi A., 2009. Automated facial expression recognition system. *In Proceedings of the International Carnahan Conference on Security Technology*, pp. 172-177.

Sigal L., Sclaroff, S., and Athitsos, V., 2000, "Estimation and prediction of evolving color distributions for skin segmentation under varying illumination", In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, pp 152-159.

Sinha P., 1994. Object recognition via image invariants: A case study. *Investigative Ophthalmology and Visual Science*, vol. 35, no. 4, pp. 1735-1740.

Solina F., Peer P., Batagelj B., Juvan S., Kovac J., "Color based face detection in the '15 seconds of face' artinstallation", In: Mirage 2003, Conference on Computer Vision / Computer Graphics  Collobration for Mode-Based Imaging, Rendering, Image Analysis and Graphical special Effects, March 10-11 2003, INRAA Rocquencourt, France Wilfried Philips, Rocquencourt, pp. 38-47, 2003.

Tao H. and Huang T.S., 1999. Explanation-based facial motion tracking using a piecewise bezier volume deformation mode. *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '99)*, vol. 1, pp. 611-617.

Teague M.R., 1980. Image analysis via the general theory of moments. *Journal of Optical Society of America*, **70**(8): pp. 920-930.

Teh C.H., Chin, R.T., 1988, "On image analysis by the methods of moments", IEEE Trans. Pattern Anal Machine Intell. 10, pp 496-513.

Terrillon J. C., Shirazi M. N., Fukamachi H. and Akamatsu S., 2000. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. *In Proc. of the International Conference on Face and Gesture Recognition*, pp. 54-61.

Vezhnevets V., Sazonov V., Andreeva A., 2003. A survey on pixel-based skin color detection techniques.  *In Proc. GRAPHICON-2003*, pp. 85-92.

Viola P. and Jones M., 2001. Rapid object detection using a boosted cascade of simple features. *In Proc. of CVPR*.

Vural E., Cetin M., Ercil A., Littlewort G., Bartlett M. and Movellan J., 2008. Automated drowsiness detection for improved driving safety. *In Proceedings of the International Conference on Automotive Technologies*.

Wang P., Ji Q., 2004. Multi-view face detection under complex scene based on combined SVMs. *Proceedings of the 17th Intl. Conf. on Pattern Recognition (ICPR 2004)*, vol. 4, pp. 179-182.

Wu B., Ai H., Huang C. and Lao S., 2004. Fast rotation invariant multi-view face detection based on real adaboost. *In Proc. of IEEE Automatic Face and Gesture Recognition*.

Xiao J., Moriyama T., Kanade T. and Cohn  J.F., 2003. Robust full - motion recovery of head by dynamic templates and re-registration techniques. *Int'l J. Imaging Systems and Technology*, vol. 13, no. 1, pp. 85-94.

Yang G. and Huang T. S., 1994. Human face detection in complex background. *Pattern Recognition*, vol. 27, no. 1,  pp. 53-63.

Yang M. H., Kriegman D. J. and Ahuja N., 2002. Detecting faces in images: A survey. *IEEE Trans. On PAMI*, **24**(1): pp. 34-58.

Zafeiriou S. and Petrou M., "Sparse representations for facial expressions recognition via l1 optimization", IEEE Conf. on Computer Vision and Pattern RecognitionWorkshops (CVPRW), 2010.

Zarit B. D., Super B. J. and Quek F. K. H., 1999. Comparison of five color models in skin pixel classification. *International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems*, pp. 58-63.

*Other Publications*


Ahlberg J. 1999. A system for face localization and facial feature extraction. *Tech. Rep. LiTH-ISY-R-2172, Linkoping University*.

Friedman J., Hastie T. and Tibshirani R., 1998. Additive logistic regression : A statistical view of boosting. *Technical report, Dept. of Statistics, Stanford University*.

Jones M., Rehg J.., 2002. Compaq skin database. http://www. crl.research. digital.com/publications/techreports/abstracts/98-11 [Accessed on 21st June 2010]

Peer P, 2003.CVL Face Dataset, http://www.lrv.fri.uni-lj.si/facedb.html [Accessed on 2nd February 2011].

Poynton C. A., 1995. Frequently asked questions about color. In ftp://www.inforamp.net/pub/users/poynton/doc/colour/ColorFAQ.ps.gz [accessed on 10th May 2010].

Zhang C., Zhang Z.,  2010. A survey of recent advances in face detection. *Tech. Rep., Microsoft Research*.

# APPENDICES

**APPENDIX 1 : Nonlinear Transformation of Chroma and Skin Model**

In $YC_rC_b$ color space, $C_r$ and $C_b$ can be defined as a function of $Y$; namely, $C_r(Y)$ and $C_b(Y)$. If the transformed chroma is expressed $C_r'(Y)$ and $C_b'(Y)$, the skin color model is specified by the centers, denoted as $\overline{C_b}(Y)$ and $\overline{C_r}(Y)$, spread of the cluster, denoted as $W_b(Y)$ and $W_r(Y)$ and is used to compute non-linear transformed chroma. See in the Figure A1.1.

**Figure A1.1: Transformed $C_r$ and $C_b$**
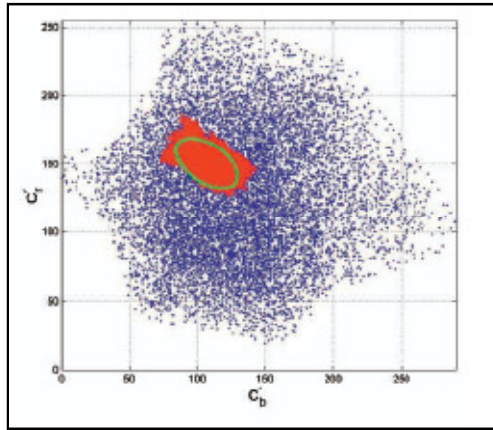


Source*:* Jain et al. )2002)

The transformation is;

$$C_i'(Y) = \begin{cases} \left(C_i(Y) - \overline{C_\iota}(Y)\right) \frac{W_{C_i}}{W_{C_i}(Y)} + \overline{C_\iota}(K_h) & if \ Y < \ K_l \ or \ K_h < Y \\ \\ C_i(Y) & if \ Y \in [K_l, K_h] \end{cases} \tag{A1.1}$$

$$W_{C_i}(Y) = \begin{cases} WL_{C_i} + \frac{(Y-Y_{min})(W_{C_i}-WL_{C_i})}{K_l-Y_{min}} & if \ Y < K_l \\ \\ WH_{C_i} + \frac{(Y_{max}-Y)(W_{C_i}-WH_{C_i})}{Y_{max}-K_h} & if \ K_h < Y \end{cases} \tag{A1.2}$$

$$\overline{C_b}(Y) = \begin{cases} 108 + \frac{(K_l-Y)(118-108)}{K_l-Y_{min}} & if \ Y < K_l \\ \\ 108 + \frac{(Y-K_h)(118-108)}{Y_{max}-K_h} & if \ K_h < Y \end{cases} \tag{A1.3}$$

$$\overline{C_r}(Y) = \begin{cases} 154 - \frac{(K_l-Y)(154-144)}{K_l-Y_{min}} & if\ Y < K_l \\ 154 + \frac{(Y-K_h)(154-132)}{Y_{max}-K_h} & if\ K_h < Y \end{cases} \tag{A1.4}$$

$C_i$, $C_b$ and $C_r$ are depicted in the previous formulas show that $W_{C_b} = 46.97$, $WL_{C_b} = 23$, $WH_{C_b} = 14$, $W_{C_r} = 38.76$, $WL_{C_r} = 20$, $WH_{C_r} = 10$, $K_l = 125$, $K_h = 188$. These parameters are estimated from HHI image dataset, and $Y_{min} = 16$ and $Y_{max} = 235$.

The elliptical model for the skin tones are described;

$$\frac{(x-ec_x)^2}{a^2} + \frac{(y-ec_y)^2}{b^2} = 1 \tag{A1.4}$$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} C_b' - c_x \\ C_r' - c_y \end{bmatrix} \tag{A1.5}$$

The parameters in the Equation A.14 and A1.5 are $c_x = 109.38$, $c_y = 152.02$, $\theta = 2.53$ radian, $ec_x = 1.6$, $ec_y = 2.41$, $a = 25.39$ and $b = 14.03$. These parameters are extracted from $C_r'(Y)$ and $C_b'(Y)$ color spaces.

# APPENDIX 2 :  Histogram of the Invariant Features

**Figure A2.1: Histogram of the geometric shape features belong to face regions. Shapes denote the feature no of the geometric shape feature.**
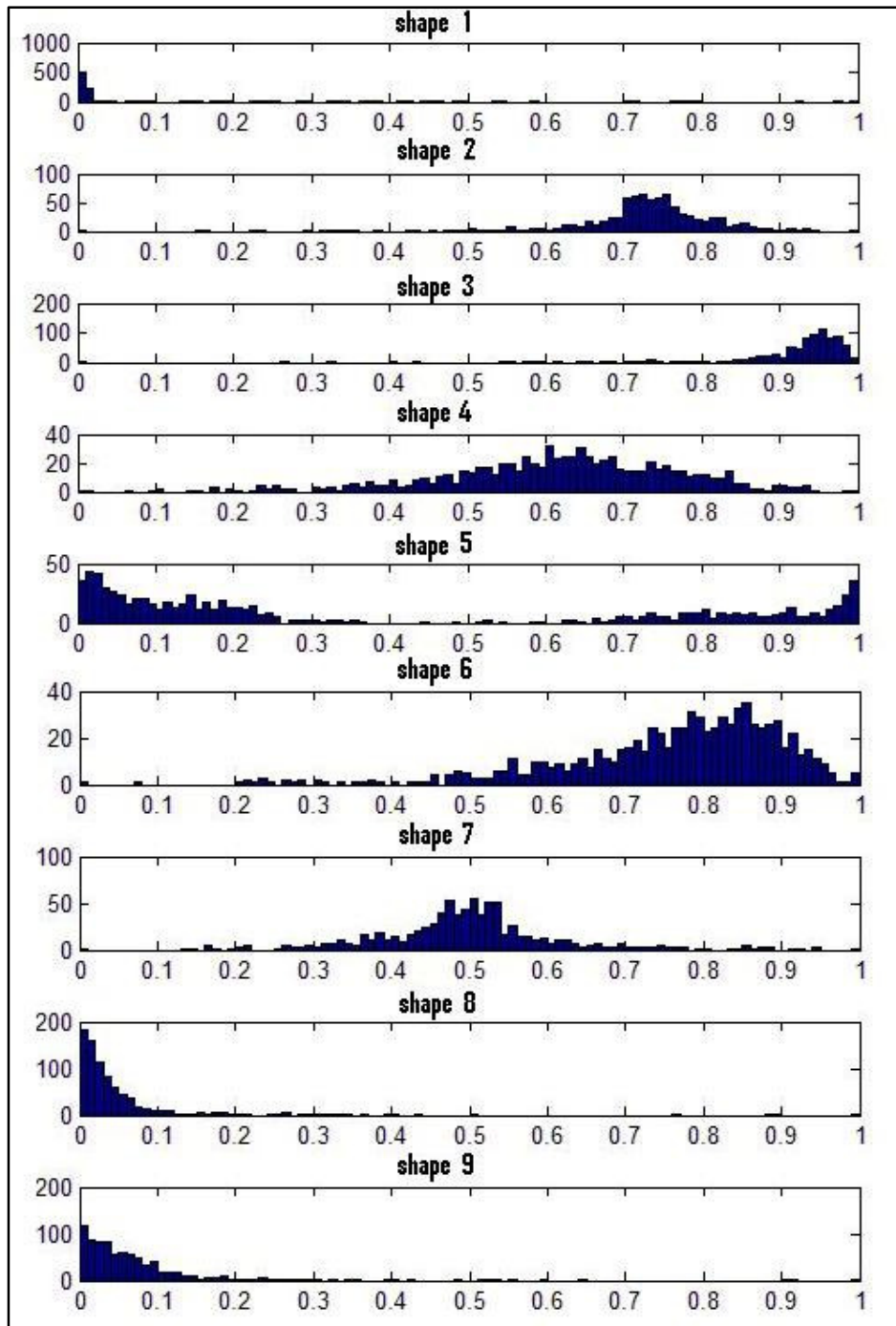
**Figure A2.2: Histogram of the geometric moment features belong to face regions. Geometric denotes the feature no of the moment.**
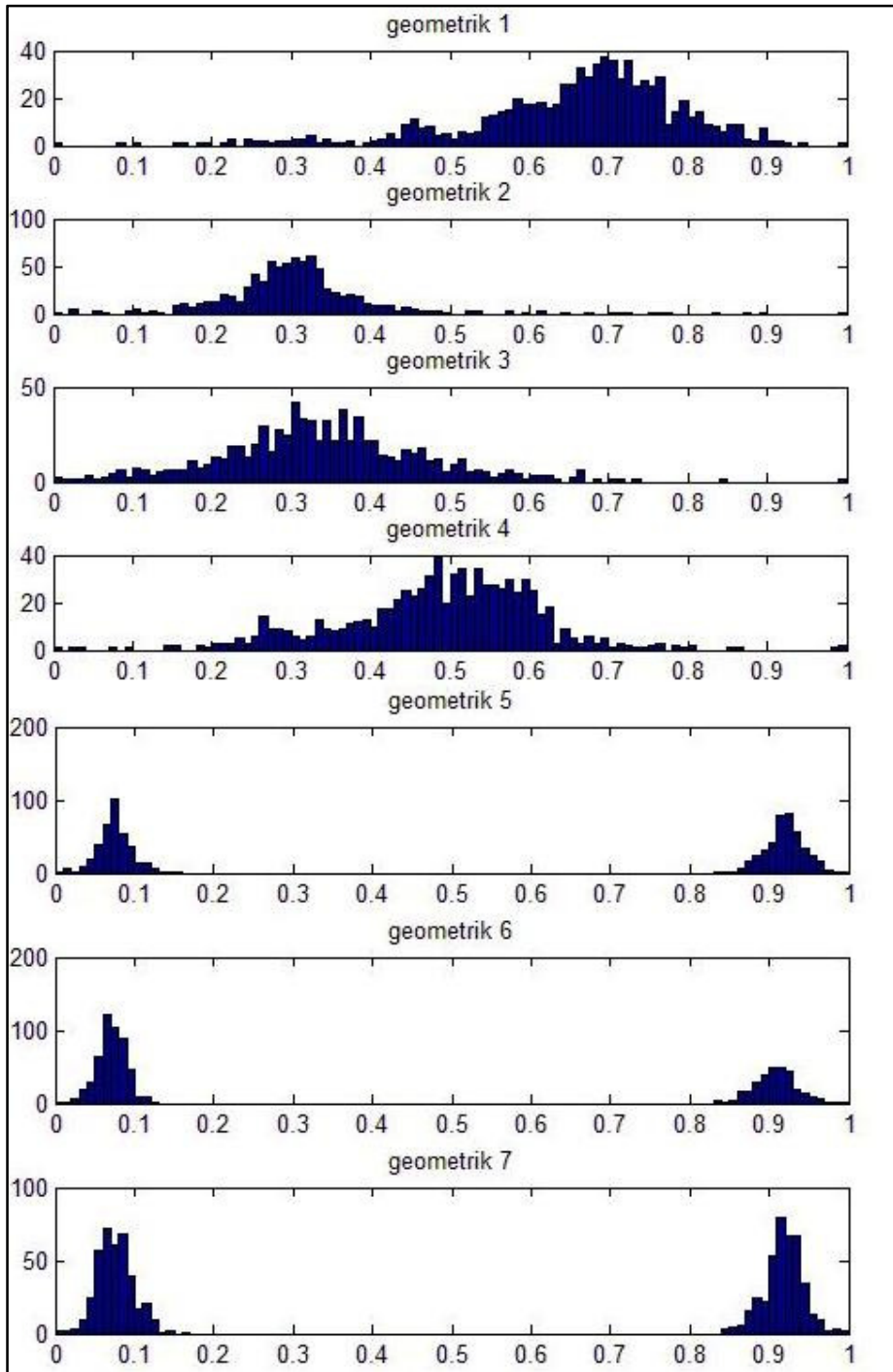
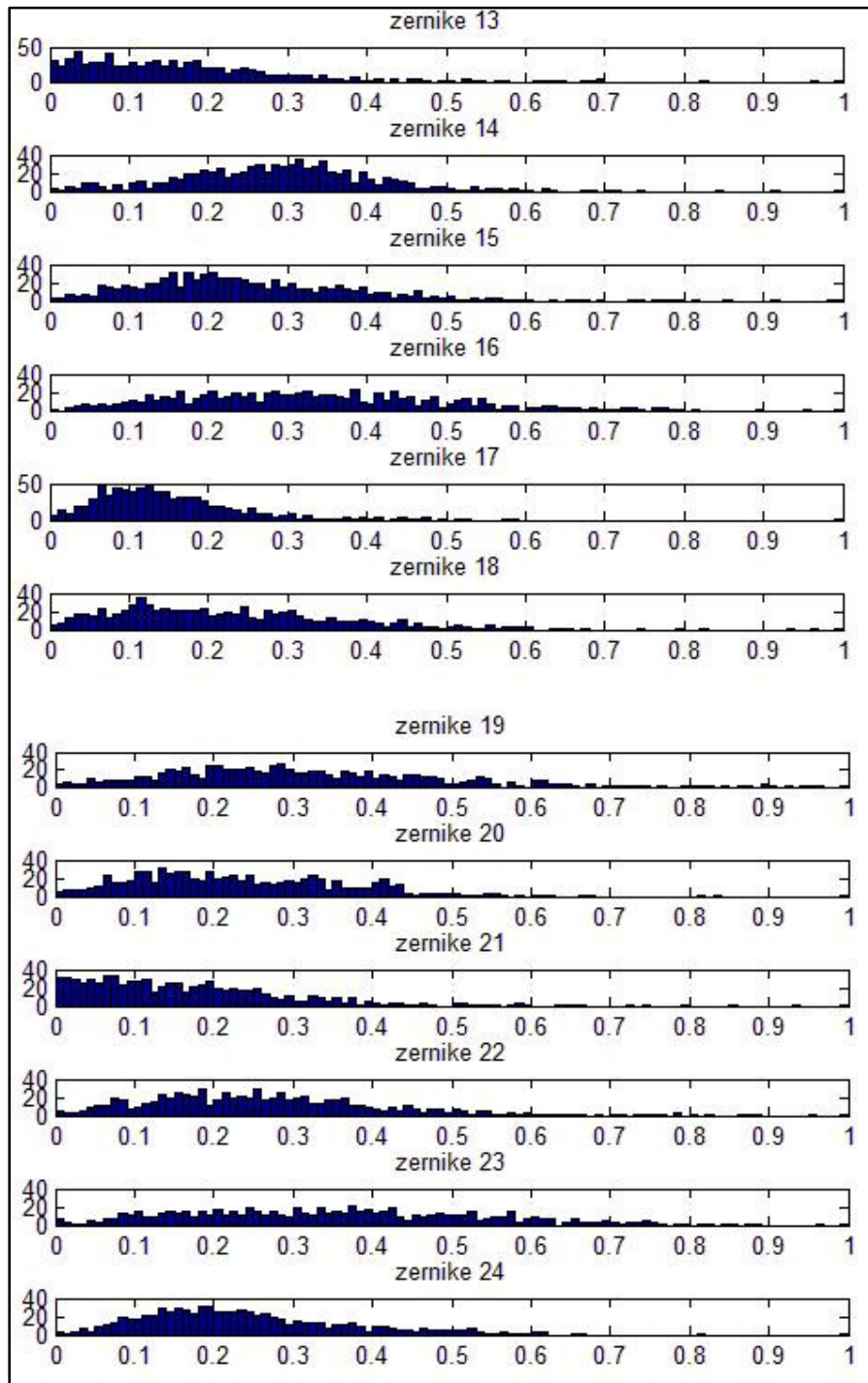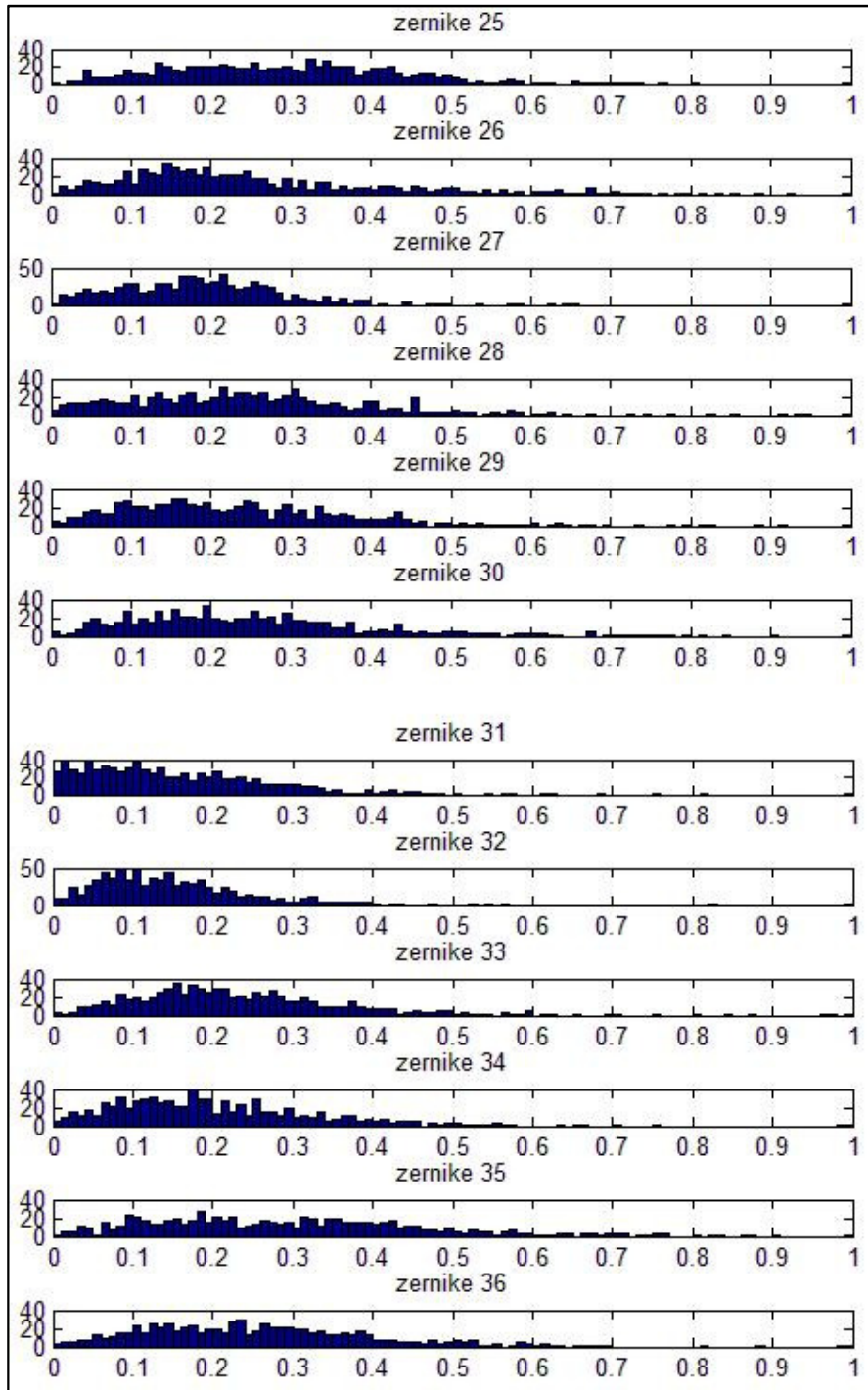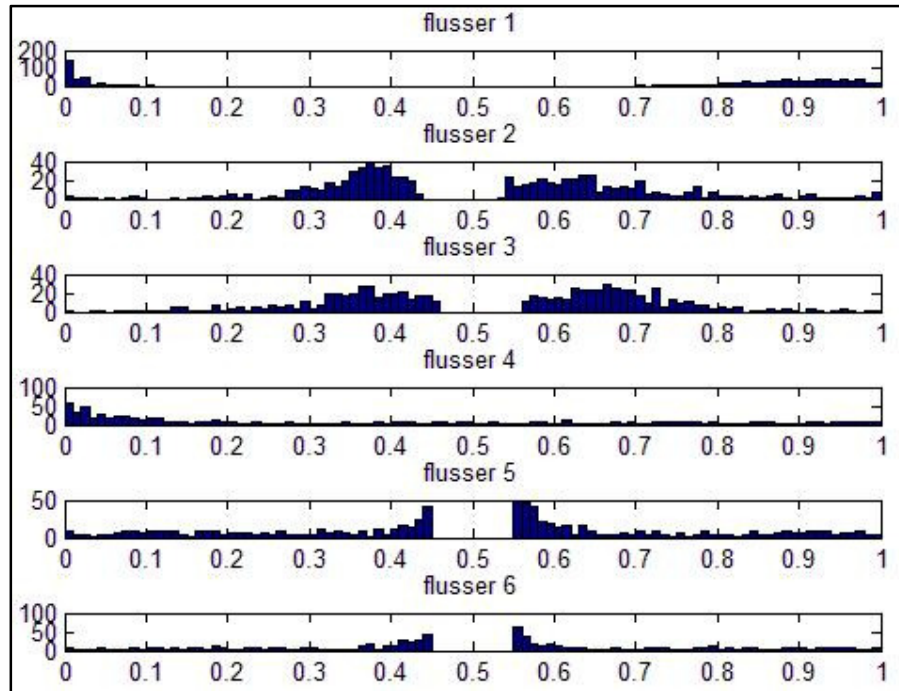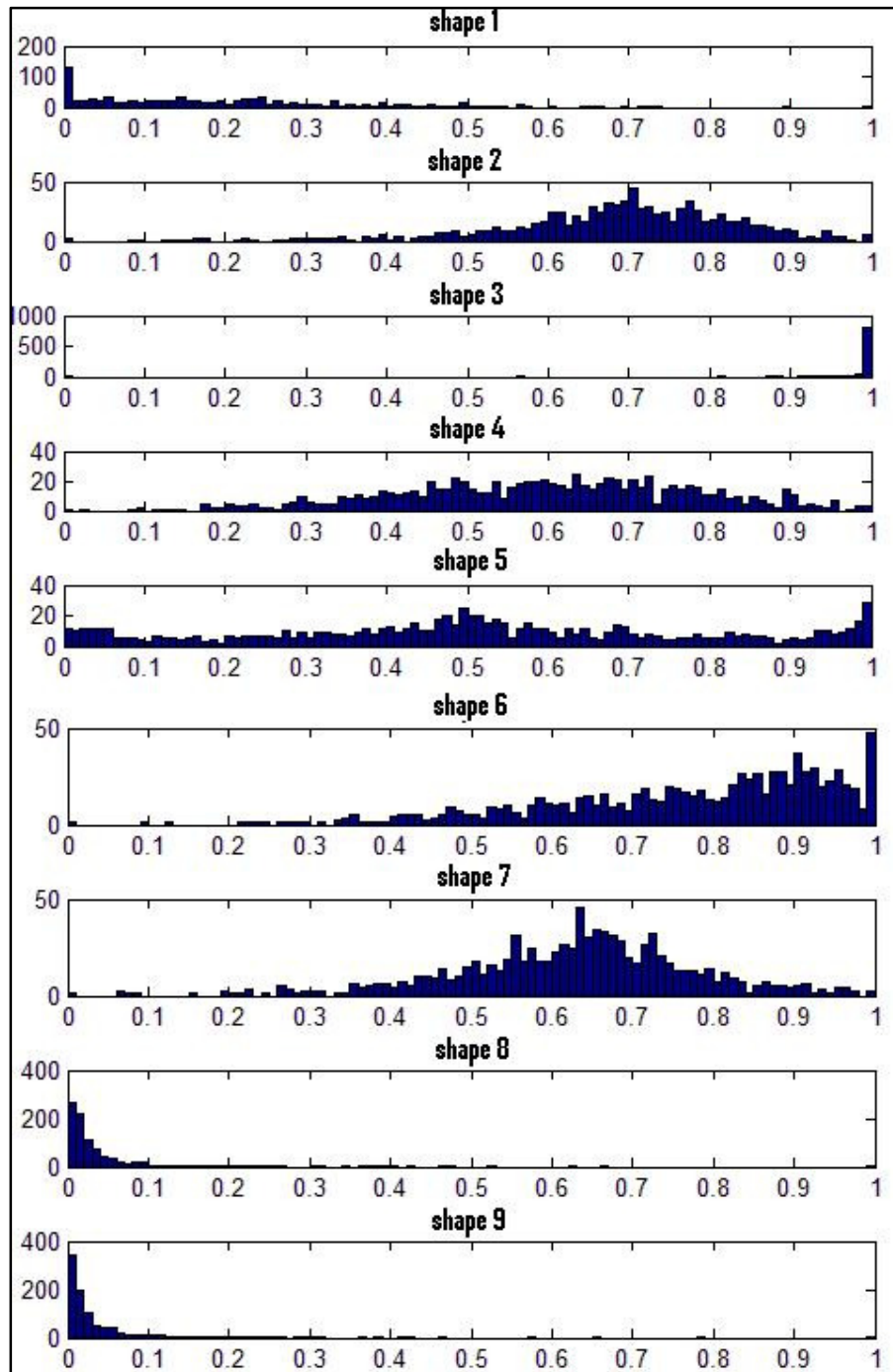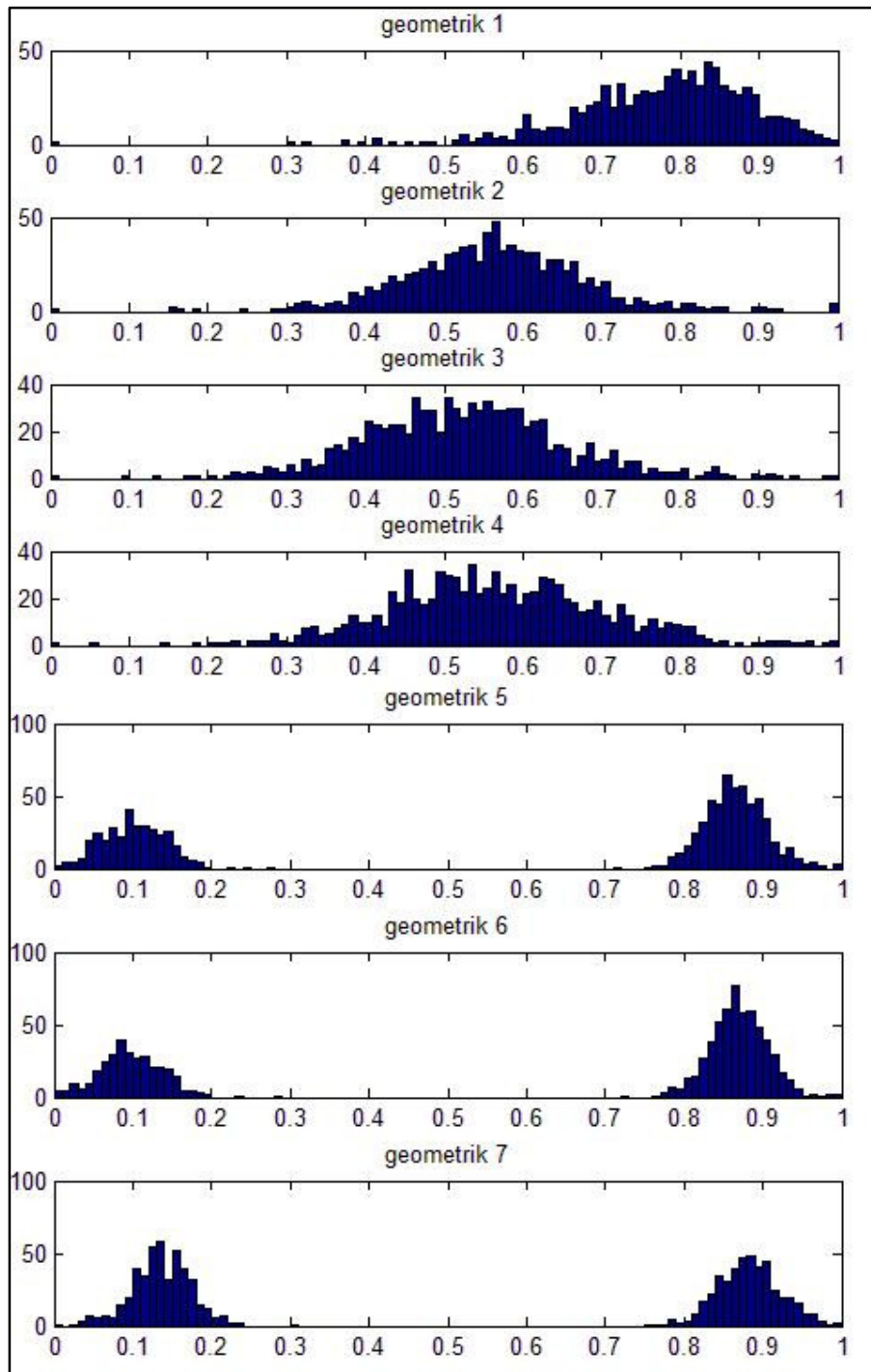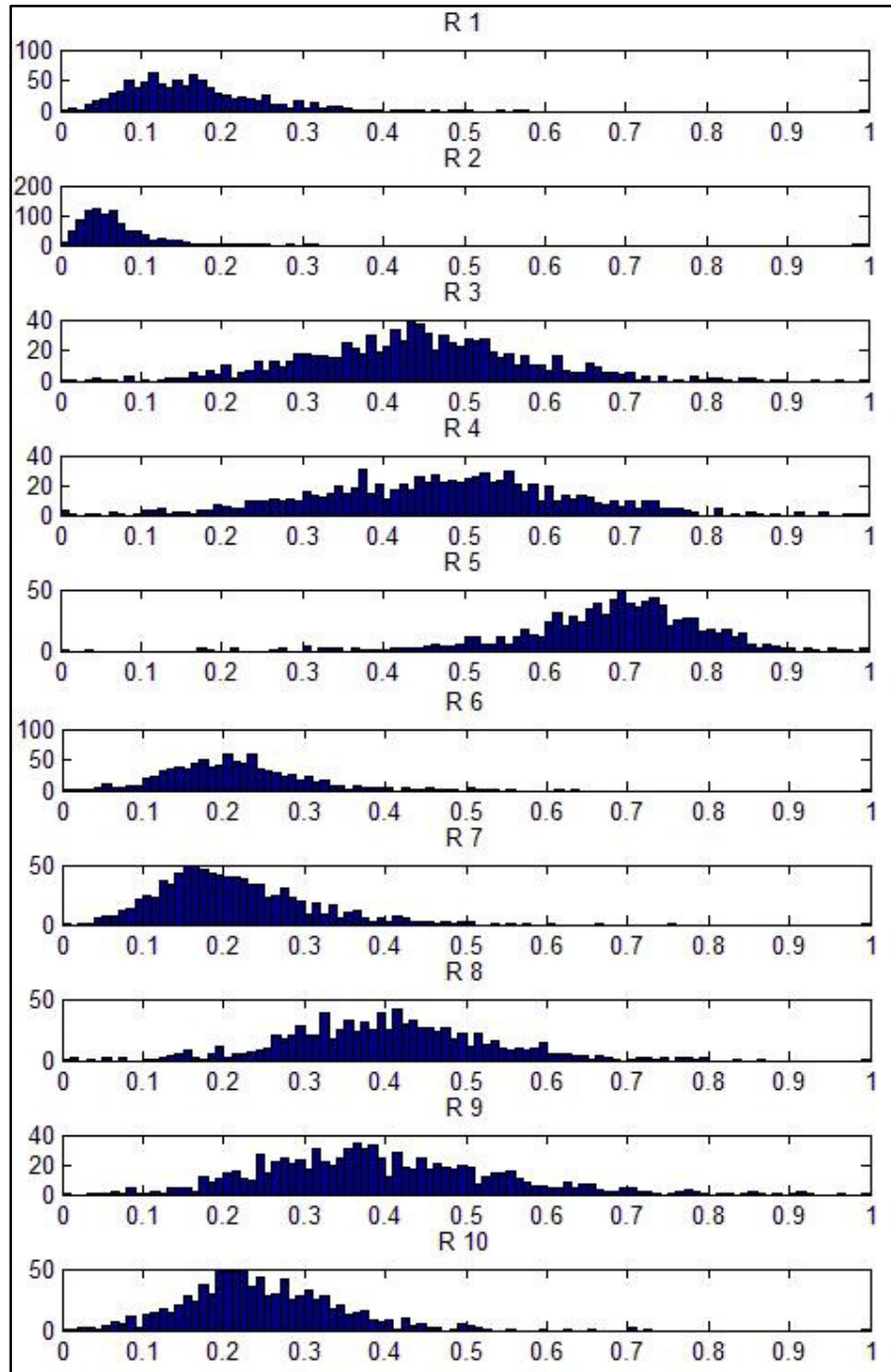**Figure A2.3: Histogram of the R moment features belong to face regions. R denotes the feature no of the moment.**
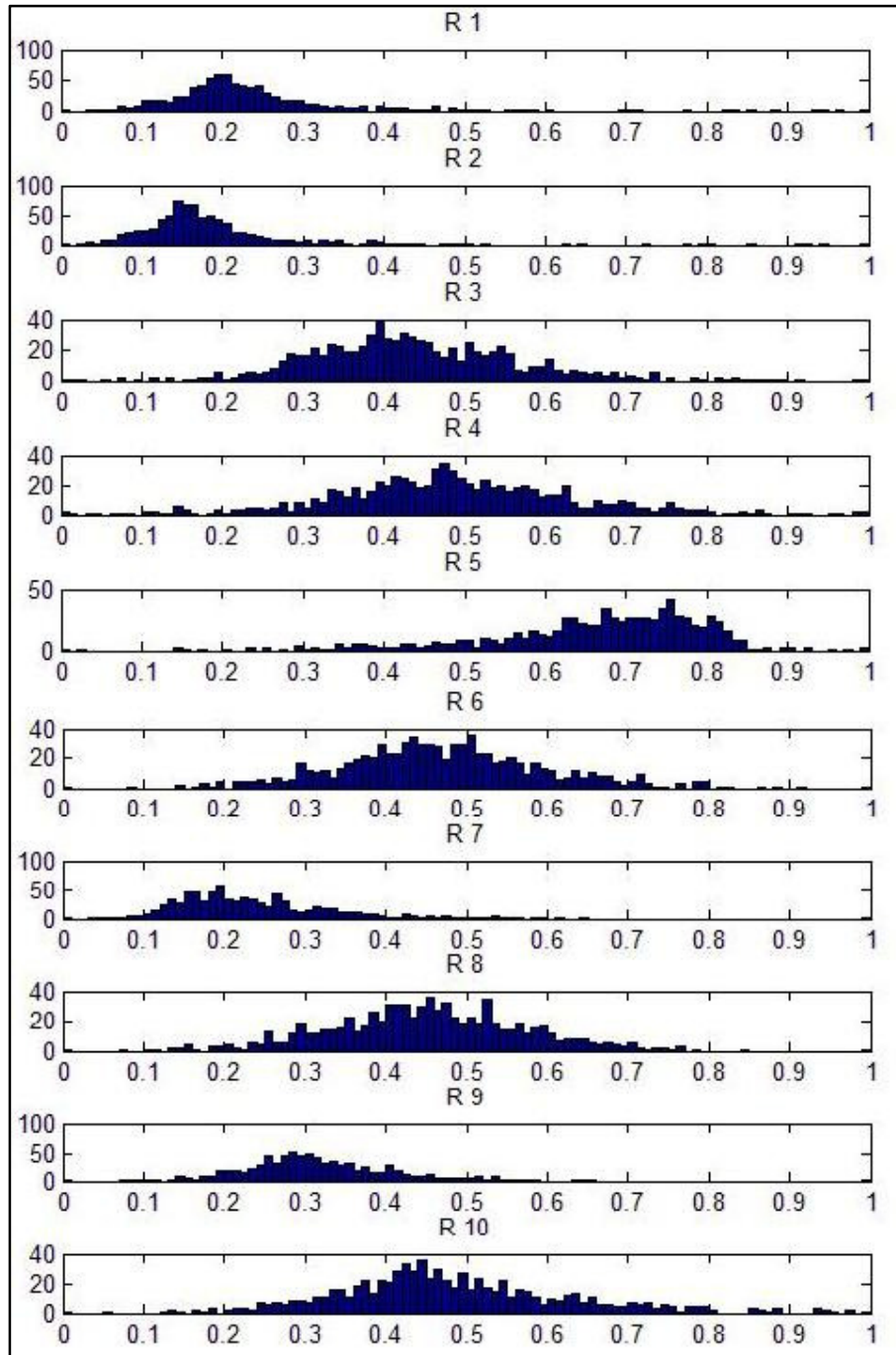
**Figure A2.4: Histogram of the Zernike moment features belong to face regions.**

**Zernike denotes the feature no of the moment 1-12.**
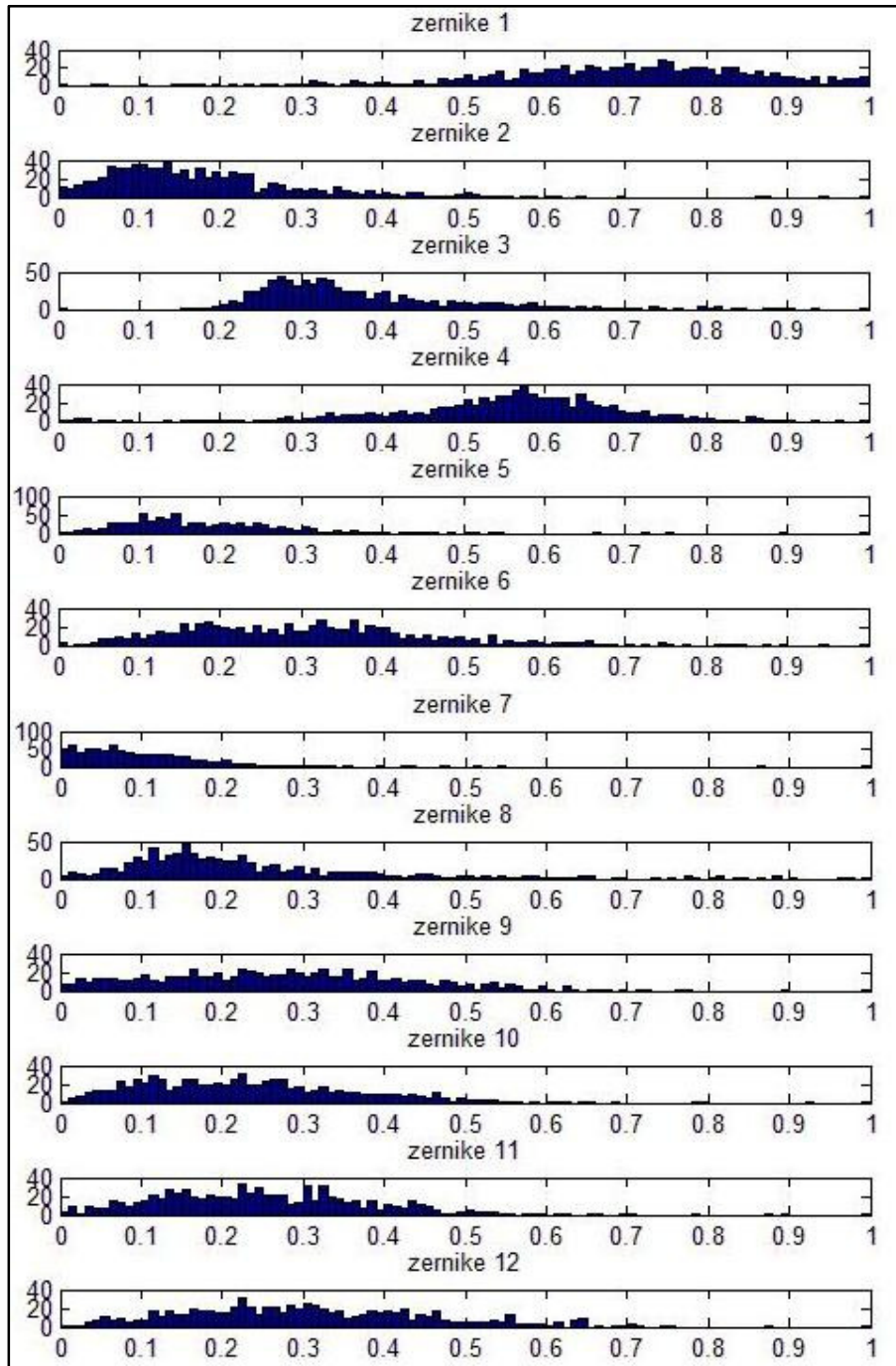
**Figure A2.5: Histogram of the Zernike moment features belong to face regions. Zernike denotes the feature no of the moment 13-24.**

**Figure A2.6: Histogram of the Zernike moment features belong to face regions.**
**Zernike denotes the feature no of the moment 25-36.**

**Figure A2.7: Histogram of the Flusser moment features belong to face regions.**

**Flusser denotes the feature no of the moment.**

**Figure A2.8: Histogram of the geometric shape features belong to non-face regions. Shapes denote the feature no of the geometric shape feature.**

**Figure A2.9: Histogram of the geometric moment features belong to non-face regions. Geometric denotes the feature no of the moment.**

**Figure A2.10: Histogram of the R moment features belong to non-face regions.**
**R denotes the feature no of the moment.**

**Figure A2.11: Histogram of the Zernike moment features belong to non-face regions. Zernike denotes the feature no of the moment 1-12.**
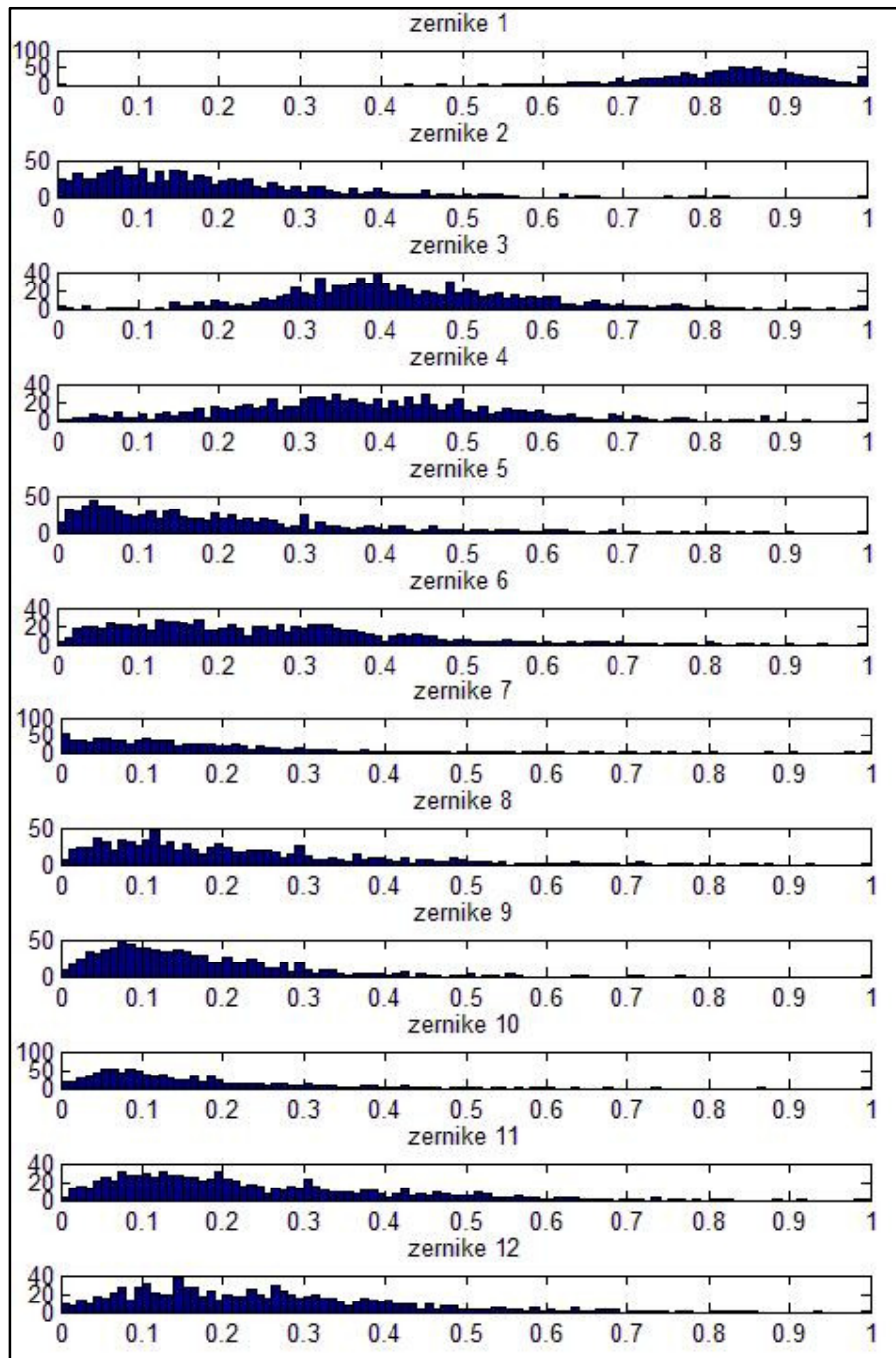
**Figure A2.12: Histogram of the Zernike moment features belong to non-face regions. Zernike denotes the feature no of the moment 13-24.**
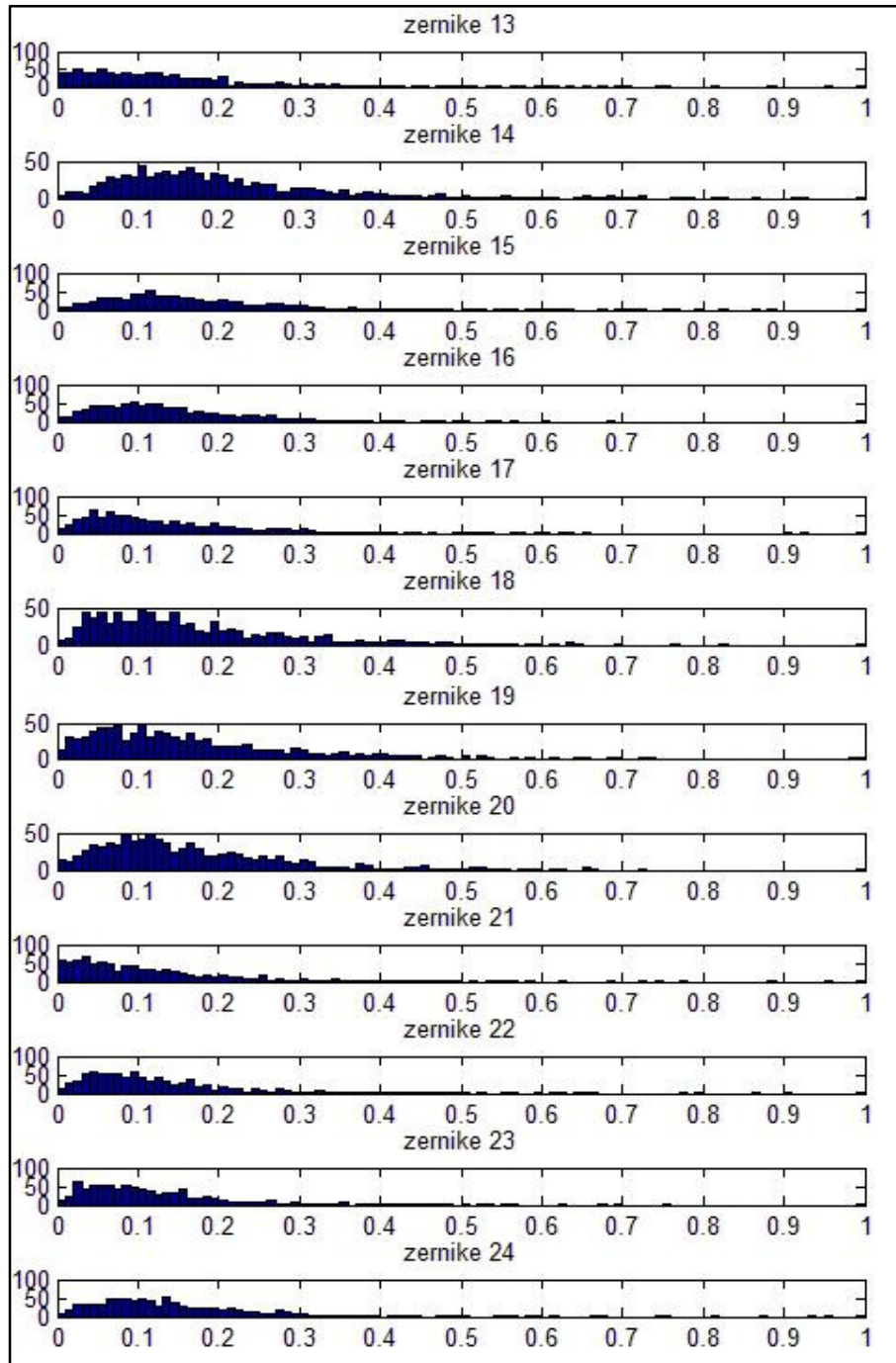
**Figure A2.13: Histogram of the Zernike moment features belong to non-face regions. Zernike denotes the feature no of the moment 25-36.**
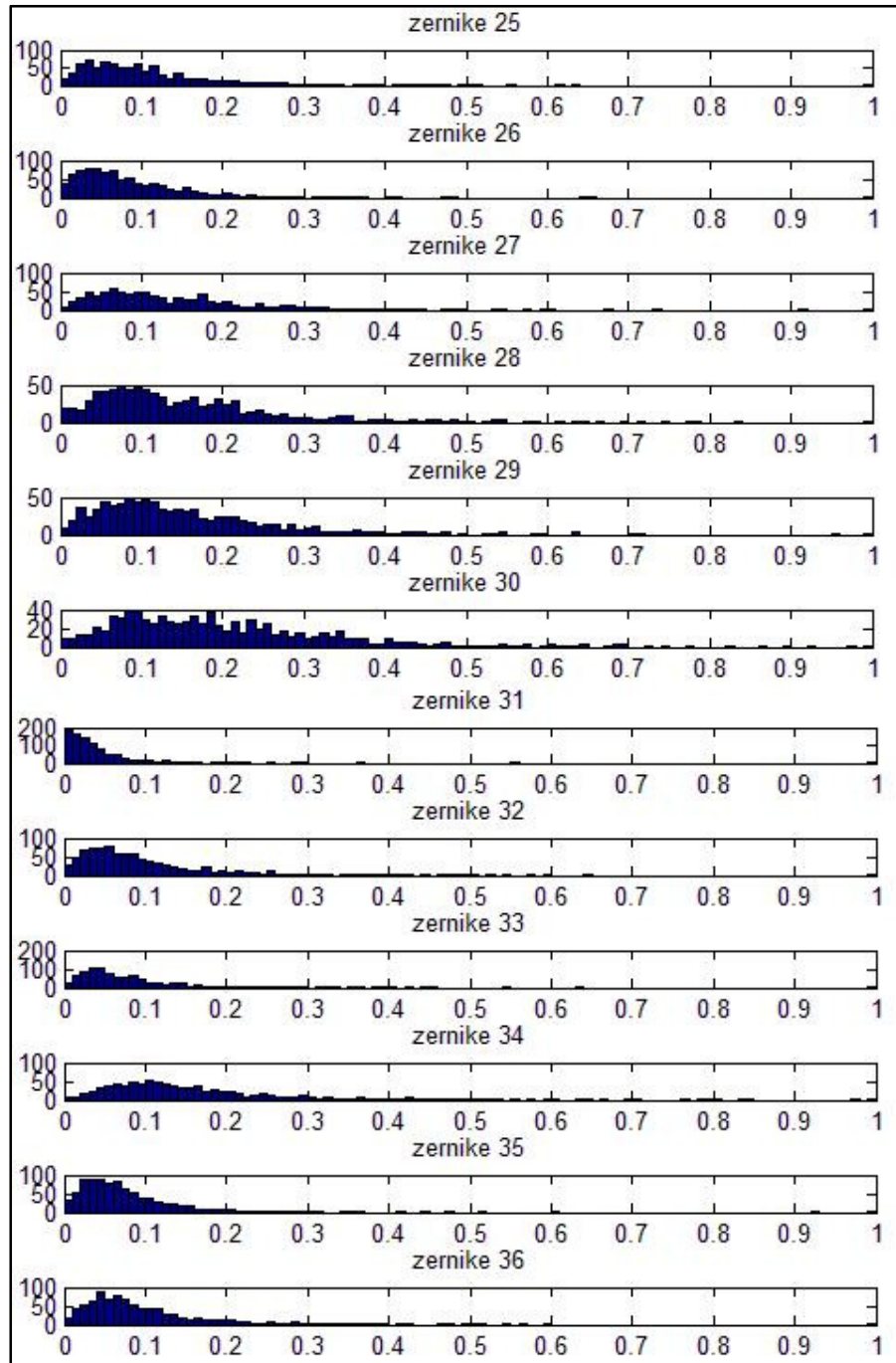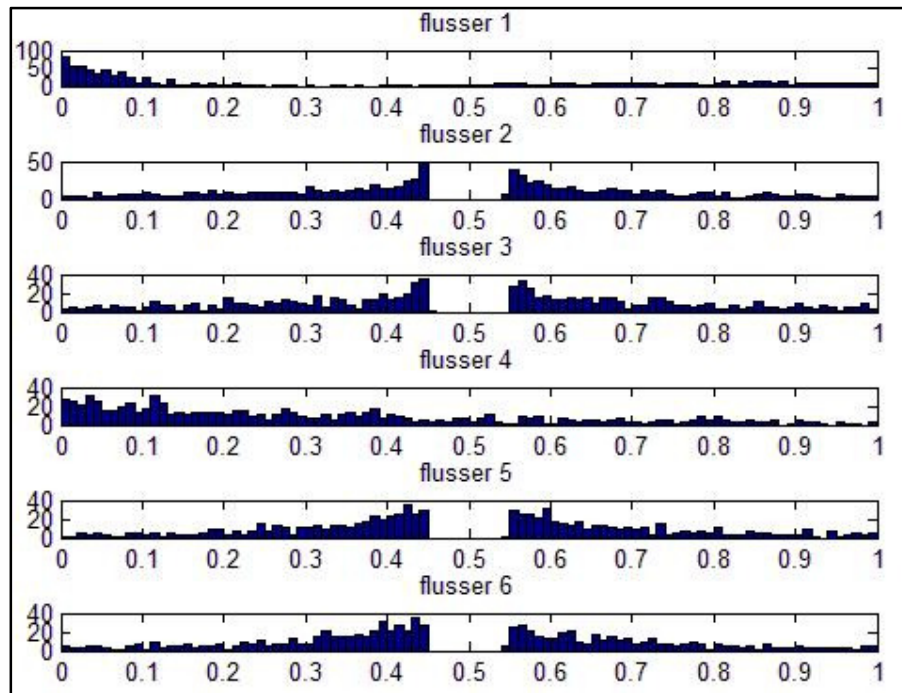
**Figure A2.14: Histogram of the Flusser moment features belong to non face regions. Flusser denotes the feature no of the moment.**

# CURRICULUM VITAE

**Name Surname :** Ali Karaali

**Permanent Address :** Osmanpaşa Mektebi Sok. No: 14 Beşiktaş-İST.

**Birthplace and Date :** İstanbul, 1985

**Second Language :** Turkish (native), English (fluent), Portuguese (basic)

**Primary School :** Alparslan  İ.Ö.O.

**Secondary School :** Rahmi Kirişçioğlu İ.Ö.O.

**High School :** Şişli Anadolu Lisesi

**Bachelor School :** Kocaeli University, Electronics and Telecommunication Engineering

**Graduate School :** Bahçeşehir University

**Institute Name :** Graduate School of Natural and Applied Sciences

**Program :** Electrical and Electronics Engineering

**Work Life :** Bahçeşehir University, Faculty of Engineering,
Electrical and Electronics Engineering Department, Research and Teaching
Assistant (February, 2009 - February, 2012)