RECOGNITION OF HUMAN FACE EXPRESSIONS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

EMRAH ENER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

SEPTEMBER 2006

Approval of the Graduate School of Natural and Applied Sciences

Prof. Dr. Canan Özgen
Director

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

Prof. Dr. İsmet Erkmen
Head of Department

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

Prof. Dr. Mete Severcan
Supervisor

Examining Committee Members

Prof. Dr. Kemal Leblebicioğlu (METU, EE) _____

Prof. Dr. Mete Severcan (METU, EE) _____

Assoc. Prof. Gözde Bozdağı Akar (METU, EE) _____

Assoc. Prof. Aydın Alatan (METU, EE) _____

Özgür Tuncer (M.Sc.) (ASELSAN) _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Emrah Ener

Signature :

ABSTRACT

RECOGNITION OF HUMAN FACE EXPRESSIONS

Ener, Emrah

MSc., Department of Electrical and Electronics Engineering

Supervisor: Prof. Dr. Mete Severcan

September 2006, 111 pages

In this study a fully automatic and scale invariant feature extractor which does not require manual initialization or special equipment is proposed. Face location and size is extracted using skin segmentation and ellipse fitting. Extracted face region is scaled to a predefined size, later upper and lower facial templates are used for feature extraction. Template localization and template parameter calculations are carried out using Principal Component Analysis. Changes in facial feature coordinates between analyzed image and neutral expression image are used for expression classification. Performances of different classifiers are evaluated. Performance of proposed feature extractor is also tested on sample video sequences. Facial features are extracted in the first frame and KLT tracker is used for tracking the extracted features. Lost features are detected using face geometry rules and they are relocated using feature extractor. As an alternative to feature based technique an available holistic method which analyses face without partitioning is implemented. Face images are filtered using Gabor filters tuned to different scales and orientations. Filtered images are combined to form Gabor jets. Dimensionality of Gabor jets is decreased using Principal Component Analysis. Performances of

different classifiers on low dimensional Gabor jets are compared. Feature based and holistic classifier performances are compared using JAFFE and AF facial expression databases.

Keywords: Facial Expression Recognition, Facial Feature Extraction, Principal Component Analysis, Skin Segmentation, Ellipse Fitting, Facial Feature Tracking, Gabor Filters

ÖZ

İNSAN YÜZ İFADELERİNİN TANINMASI

Ener, Emrah
Yüksek Lisans, Elektrik Elektronik Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Mete Severcan

Eylül 2006, 111 Sayfa

Bu çalışmada tam otomatik ve ölçeklendirmeden bağımsız bir nite çıkarıcı önerilmektedir. Yüzün konumu ve boyutları ten bölütlemesi ve elipse oturtma ile çıkarılmaktadır. Ayrıştırılan yüz bölgesi daha önceden tanımlı bir boyuta getirilmekte, alt ve üst yüz şablonları kullanılarak yüz niteleri ayrıştırılmaktadır. Şablon lokalizasyonu ve şablon parametrelerinin hesaplanması işlemleri ana bileşen analizi yöntemi ile yapılır. Nötr ifadeli resim ve analiz edilen resimler arasında yüz nitelerinin konumlarının değişimi resmin ifadeye göre sınıflandırılmasında kullanılır. Değişik sınıflandırıcıların performansları değerlendirilmiştir. Önerilen nite ayrıştırıcının performansı örnek video dizileri üzerinde de test edilmiştir. Yüz niteleri ilk film karesinde ayrıştırılmış ve KLT izleyici ile bu ayrıştırılan niteler izlenmiştir. Kaybolan niteler yüz geometrisi kuralları kullanılarak belirlenmiş ve nite ayrıştırıcı yardımıyla tekrar konumlandırılmıştır. Nite temelli tekniğe alternatif olarak yüzü parçalara ayırmadan analiz eden bir bütünsel teknik gerçeklenmiştir. Yüz resimleri değişik ölçek ve yönlendirmelere ayarlanmış Gabor filtrelerce filtrelenmiştir. Filtrelenmiş resimler Gabor jetleri oluşturmak üzere birleştirilmiştir. Gabor jet boyutları ana bileşen analizi yöntemi kullanılarak azaltılmıştır. Boyutu

düşürülen Gabor jetleri üzerinde farklı sınıflandırıcıların performansları karşılaştırılmıştır. Nite tabanlı ve bütünsel sınıflandırıcıların performansları JAFFE ve AF yüz ifadesi veri tabanları kullanılarak test edilmiştir.

Anahtar Kelimeler: Yüz İfadelerinin Tanınması, Yüz Nitelerinin Ayrıştırılması, Ana Bileşen Analizi, Ten Bölütlemesi, Elipse Oturtma, Yüz Nitelerinin İzlenmesi, Gabor Filtreleri

To My Parents & Aysun

ACKNOWLEDGEMENTS

I want to express sincere thanks to my supervisor, Prof. Dr. Mete Severcan for his valuable guidance, endless patience and helpful suggestions throughout this study.

I would like to thank to Aysun and my family for their valuable support, trust and encouragement during this thesis.

I also would like to thank to ASELSAN INC especially to my department for letting me to involve in this thesis work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AF	: Aleix Face
CDM	: Cosine Distance Measure
COE	: Center of Eyes
COM	: Center of Mouth
CV	: Coordinate Variance
DFT	: Discrete Fourier Transform
DSTFT	: Discrete Short Time Fourier Transform
EDM	: Euclidian Distance Measure
FA	: Facial Action
FAFE	: Fully Automatic Feature Extractor
FER	: Facial Expression Recognition
FFT	: Fast Fourier Transform
HMI	: Human Machine Interaction
ICA	: Independent Component Analysis
JAFFE	: Japanese Female Facial Expression
JTFR	: Joint Time Frequency Resolution
KLT	: Kanade Lucas Tomasi
KNN	: K Nearest Neighbour
LED	: Light Emitting Diode
LFR	: Lower Facial Region
MDM	: Mahalanobis Distance Measure
NM	: Nearest Mean
NN	: Nearest Neighbour
PCA	: Principal Component Analysis
PCC	: Principal Component Count
RF	: Receptive Field
SVM	: Support Vector Machine
UFR	: Upper Facial Region

CHAPTER 1

INTRODUCTION

As machines and people begin to co-exist and cooperatively share a variety of tasks, the need for effective communication channels between machines and humans becomes increasingly important. Systems to form these communication channels are known as human machine interaction (HMI) systems.

Advances in technology lead to the development of more effective HMI systems which no longer rely on common devices such as keyboard, mouse and displays but take commands directly from user's voice and mimics. Such systems aim to simulate human-human interaction by only using communication channels used between humans and not requiring artificial equipment.

Human-machine interaction should be enhanced to more closely simulate human-human interaction before machines take more places in our lives. Since change of expressions on human face is a powerful way of conveying emotions, facial expression recognition (FER) will be one of the greatest steps for enhancing HMI systems.

Although humans can detect and interpret expression in a scene with little or no effort, same task will be quite challenging for machines. Changes on face and body should be modeled using properly chosen features and those features should be tracked, classified in real time.

FER can be thought as an inter-disciplinary problem of image-video processing, pattern recognition, psychology and studies to increase accuracy and speed has been carried out for the last 20 years.

An important problem with recognition task is the number of expressions apart from common ones such as anger, joy and fear. Several other messages should be recognized in such a system. A smiling mouth and raised eyebrows meaning "I don't know" can be given as an example to such messages. There are also expressions commonly used by a specific person but very rare in public. This makes the problem specific for each person and requires an adaptive classification mechanism in interpretation of subject's expression.

Apart from HMI systems video telephony and conferencing is another application of facial feature extraction. High bandwidth required to transmit facial images brings the idea of using an avatar which simulates changes in users face. Changes on facial feature coordinates can be coded and transmitted on low bandwidth channels enabling video conferencing and video telephony while protecting user's privacy.

1.1.System Overview and Outline of Thesis

In this study two methods (holistic and feature based) are proposed to classify images according to the expression. Their performances are compared using two facial expression databases. Feature based method is based on extraction of key facial features corresponding to eyes, eyebrows and mouth. A technique named as FAFE (Fully Automatic Feature Extractor) is proposed to extract these key features without any manual initialization.

Feature based method proposed in this study can be divided into two steps:

1.) Finding location, size and orientation of face in still image and extraction of facial features using FAFE.

2.) Classification of image according to expression using extracted features.

Face localization algorithm used in this work starts with skin segmentation to locate rough coordinate of the face. Algorithm is designed to work with images containing over shoulder view of a single person. With the help of this constraint, best ellipse found on the extracted skin image is assumed to be the face region.

Feature extraction step of FAFE starts with locating the center point between eyes and eyebrows named as COE (Center of Eyes).After locating COE in the image 14 facial features defining the size and location of eyes and eyebrows are extracted using template matching.

A search window for mouth region is defined by using the distance between eyes. In this region the center of mouth (COM) is found and 4 feature points belonging to mouth are extracted. Distances between predefined feature point set (feature distances) are calculated and used for classification.

Feature based classifier requires the availability of neutral expression view of the person whose picture will be classified according to expression. Relative position changes in the classified image are calculated using this neutral image. Although databases used to test this algorithm consist of images with an approximately constant distance to camera there are still some variations. These variations are taken into account and position changes are normalized using the distance between centers of eyes. Bayes, nearest mean (NM), nearest neighbor (NN) and K-nearest neighbor (KNN) classifiers are used to analyze movement of facial features and performances for those classifiers are compared.

As a holistic method for facial expression classification Gabor wavelet filters, which allow analysis of patterns at multiple scales and multiple orientations are used. Face region is extracted and inner face region containing eyebrows, eyes and mouth is partitioned from target image. Gabor filters tuned to different scales, different frequencies are applied and responses are collected into a matrix known as jet.

Performances of different distance measures like cosine, Euclidian and Mahalanobis distances together with NM, KNN classifiers are compared. Jet size for even a small image is quite large when large number of scales and orientations are used in Gabor filters. This makes the classification calculations time consuming. Dimensionality of the jet vector is decreased using Principal Component Analysis (PCA) and changes in classification success rate for changing number of principal components are analyzed.

Two facial expression databases known as Japanese Female Facial Expression Database (JAFFE) and Aleix face database (AF) are used to test both algorithms.
In Chapter 2 face localization techniques in literature and face localization subsystem of this study will be described. In Chapter 3 algorithms to define facial feature search windows and feature extractor implemented in this study are described. Some alternative methods for center of eye localization together with their shortcomings are presented. A facial feature tracker using FAFE and KLT (Kanade-Lucas-Tomasi) tracker is also proposed in this chapter. In Chapter 4 information about Gabor filters will be presented and feature extraction using those filters will be described. Performances of proposed holistic and feature based expression analysis techniques and parameters effecting their success rate are presented in Chapter 5. Results obtained using two different facial image databases are also included in this chapter. In Chapter 6, conclusions on this study are made.

CHAPTER 2

FACE LOCALIZATION

A common problem in face analysis studies is detection and localization of face region on images. Several techniques have been proposed to locate faces in images. First difficulties with face localization and a classification for available face localization techniques will be described (section 2.1). Face localization technique used in this study (as a part of FAFE) which is based on skin segmentation and ellipse fitting will be described in section 2.2.

2.1.An Introduction to Face Localization

The first stage of a fully automatic facial feature extractor is locating the face region and its boundaries. A complete face localization system should cope with several challenges some of which are described below.

Pose: Facial features including eyes, nose, and mouth may partially get invisible or distorted because of the relative pose of face and camera.

Occlusion: Facial features may get occluded by beard, mustache, and glasses. Similarly make-up can cause appearance of artificial regions on face or hide normal facial boundaries.

Expression: Facial features show great changes in their shape under different expressions. Some features may get invisible or some only get visible under different expressions.

Imaging Conditions: Lighting and changes in camera characteristics significantly effect chrominance of face regions. Some features may be occluded or get combined with shadows or shines on face and cause color information loss.

Several techniques for locating faces in images are proposed. Although a strict classification of these algorithms is not possible because of their highly overlapping boundaries, a grouping to these algorithms is proposed in [7]. Face localization techniques can be divided into four main categories which are given in Figure 2.1

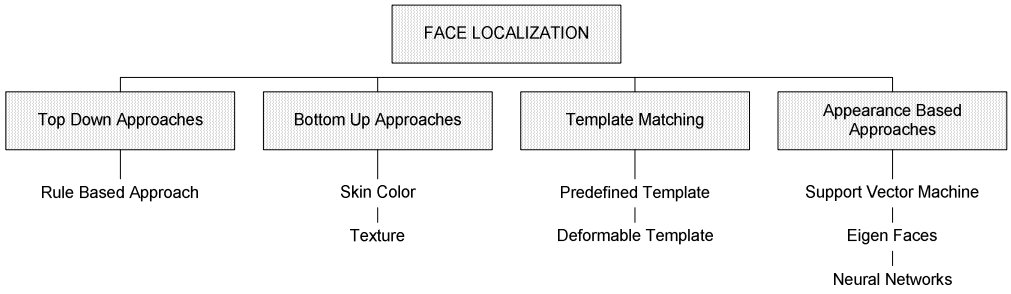


Figure 2-1 Categorization of Face Localization Techniques

2.1.1.Top Down Approaches

It is possible to define some rules which define relations between features of a face. For example a face includes two eyes one nose and mouth with clearly defined relative positions between each other. A top down approach uses hierarchical rules. Most general rules were checked in lowest resolution images while most detailed rules tested in highest level images.

Yang and Huang [24] proposed an early example of top-down face localization. At the lowest level they modeled face as having four neighbor cells with approximately equal intensity and cells surrounding them with a significantly lower intensity. At the highest resolution images, facial features extracted by edge detection are tested for acceptable configurations between each other.

Another top down approach was proposed by Kotropoulos and Pitas [28]. They used horizontal and vertical intensity histograms of images to locate possible positions of face. Extracted face region histograms are later analyzed to locate facial features using abrupt changes in the vertical and horizontal histograms.

Converting human knowledge about face into rules is a challenging problem. If rules are too detailed valid face regions will be removed and if the rules are too general they may lead to false localizations. Another problem with rule based validation is to extend those rules into several pose of the face.

Face class has significant variations caused by age, orientation, illumination and skin color. These variations can be decreased by image preprocessing. Gabor wavelets and Gaussian derivative filters can be applied to yield more silent representations [7]. Shade correction filters and histogram equalization can be used to reduce variations due to illumination.

In top down approaches the amount of data required to be processed is usually too large. In order to make this approach computationally viable, the dimensionality of the input image must be kept low, however even if we restrict our search with the interior of the face, input data size would be relatively high.

2.1.2.Bottom Up Approaches

In contrast to holistic approaches bottom up approaches aim to detect separate facial features and classify plausible configurations amongst the detected features as face. Humans can effortlessly detect faces under several pose, orientation and lighting conditions so there must exist some properties of face invariant of those conditions [7]. Feature based approaches aim to find and use those properties of face. Statistical relations between the extracted regions are later used to decide on the existence and location of the face.

The main problem with this approach is that it requires an eye detector, a nose detector and a mouth detector. Problem of detecting faces has been replaced by detecting multiple deformable parts which requires higher resolution images and more computational power.

This approach also has weaknesses to cope with occlusions due to orientation, facial hair, glasses and shadows where detection of features may become

impossible. Although feature based methods have several limitations they are widely used in face detection and face tracking applications.

2.1.2.1.Skin Color

Human skin is an important feature of face. Skin extraction is proven to be a reliable feature for detection with its common usage from hand tracking to face detection. Although skin color for different people vary studies has shown the difference is mostly in their intensity rather than their chrominance.

Several color spaces and distribution models for skin color are proposed. P.Gejgus and M.Sperka [29] proposed statistical techniques for skin segmentation. They modeled skin color as a Gaussian mixture distribution in pure color space (r,g)space. Gaussian parameters for this distribution are calculated using expectation maximization algorithm and maximum likelihood technique is used for skin segmentation once Gaussian mixture is learned. S. Spors and R. Rabenstein [30] proposed a Bayes classification technique in [YCrCb] space. After extracting skin region they calculated projections of resultant binary image in x and y directions. Face is localized using means and standard deviations of extracted histograms.

2.1.2.2.Facial Features

Locating Facial features like eyes, nose, mouth directly and later extracting face boundary is another approach used in face localization. H.P.Graf [26] proposed a method to extract facial features in gray level images. Band pass filtering followed by morphological operations is applied to enhance regions with specific shape like facial organs. Connected components are chosen as facial features and their relation are analyzed to detect location of face

Eye localization is useful in applications such as face normalization, gaze-based human-computer interfaces and security systems [10]. Eyes have several physiological properties which make detection and tracking possible. For example eye blinking is a continuous and discriminative property of eyes. T.N.Bhaskar

proposed an optical flow technique coupled with frame differencing to detect eye blinking and have reported a 22 frame per second eye localizer in their study [10].

A very well known and widely used property of eyes is the red eye effect. This effect is widely used in recent studies for real time tracking of eyes. Systems utilizing this property use a black and white camera surrounded by two rings of infrared LEDS. One of those rings is along the camera axis and the other is off the axis. Camera captures two sets of interlaced images which are illuminated with one of the two rings. Those captures illuminated with LEDs on the camera axis (in axis images) include bright pupils while those illuminated with the other ring (out axis images) includes dark pupils. This phenomenon is used for direct detection of eyes. Regions except eyes are similarly illuminated in both image sets. Eyes are significantly more illuminated by in axis LEDs. Differencing of two captured images will have peaks at eye locations and thresholding is used to extract eye locations.

Although hardware for image acquisition is similar in infrared based techniques, algorithms used to improve tracking vary. A.Haro combined their system with a Kalman tracking algorithm [11]. A.Kapoor and W.Picard combined adaptive thresholding; rule based false detection removal and a multi state tracker to increase speed and accuracy of their tracker [1]. The camera in [11] is given in Figure 2.2

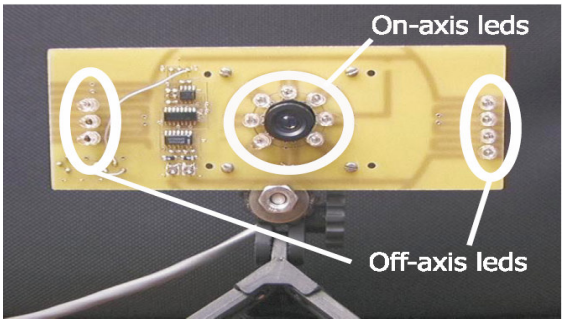


Figure 2-2 Infrared Camera for Eye Localization

2.1.2.3.Texture

Texture can be defined as images, pattern which has some structure, rules for creation. Although several variations in details exist, face can be thought as a highly defined pattern with its characteristics like symmetry and content. With the help of these characteristics face can be thought as a texture.

Studies of M.F.Augusteijn can be given as an example to texture segmentation based face detection. M.F.Augusteijn [25] defined three types of face patterns defining skin hair and other regions of face. Detection of face is converted to detection of those predefined textures and extracted regions are validated using cascade neural networks. An important shortcoming of this approach is its weakness to cope with scale.

Although partitioning face into three structures can be effective in low resolution images this partitioning will not be representative in higher resolution images which contain higher degree of detail.

2.1.3.Template Matching

In template matching the idea is to model face using a function named as template. Correlation values between template and image is calculated and presence of face is determined using those correlation values.

An important weakness of template matching in face localization is its dependence on pose variance. Although single templates can not represent face in different pose, multiple templates have shown to be representative for multiple pose face images. Face templates can have predefined shape or their shape can be deformed by changing its parameters (deformable).

2.1.3.1.Predefined Templates

Predefined templates are not deformed according to some parameter but have constant forms. Presence of faces at multiple scales can be detected by up sampling or down sampling analyzed image.

I.Craw [37] proposed a frontal view face template based approach to face localization. Edges on images are extracted using Sobel filters and those edges are grouped together using special constraints. A complete face template is used for extraction of face contour and smaller templates are used for the detection of facial features like eyes, mouth.

J.Miao [38] proposed a multiresolution multi orientation template matching procedure for face detection. Input image is rotated with angles between -20 and 20 degrees. Multiple resolution images are formed using those rotated images. Edges on images are extracted using Laplacian operators and correlation between images and face template is calculated to detect the existence of a face.

2.1.3.2.Deformable Templates

Deformable template is a priori elastic model for target object. Templates are specified by a set of parameters which can be changed to make template flexible enough to fit the searched object. Deformable template interacts with the image in a dynamic manner. An energy function is defined to link edges, peaks and valleys in input image to corresponding parameters in the template [7]. Best fit to target image is done by minimizing this energy function. Initial values for template parameters are most commonly chosen to represent the most expected shape of the template and they are modified using some update algorithm such as steepest descent to minimize energy function.

Y.H.Kwon [27] used blurring filters and morphological operators to enhance edge information extracted from input image. Hough transform is applied to find dominant ellipse shapes in image. Most probable four ellipses are analyzed using deformable templates and features in those ellipses are extracted. Relations between those features are used to validate the existence of a face in the extracted ellipses.

2.1.4.Appearance Based Methods

Template matching approach is based on templates predefined by human. In contrast to template matching based approaches appearance based approaches learn templates from training images. Appearance based approaches uses machine

learning and statistical techniques to model face. Learned models are usually in the form of distributions and discriminant functions and these models are used to distinguish between face and non-face objects.

Appearance based face localization can be thought in a probabilistic framework. Feature vector derived from an image forms a random variable which will later be classified as face or non-face. A straightforward approach is to use maximum likelihood or Bayes classification but the dimensionality of the feature vector makes this approach computationally infeasible.

Another approach to discriminate between face and non face classes is to decrease the dimensionality of feature vector by projecting it to a lower dimensional space. Later linear discriminant functions can be used to define linear boundaries between face and non-face classes. Neural networks can also be used to define non linear boundaries between classes.

2.1.4.1.Eigen Faces

Karhunen-Loeve expansion which is also known as the principal component analysis in statistical theory has wide applications in the analysis of signals and functions. L.Sirovich and M.Kirby [8] proposed an application of this method to face images. They have shown that an eigen face set of 40 images can represent an image with 3% error.

Turk and Pentland [2] extended results of [8] and showed that coordinates of a face image in eigenface subspace is an important feature for face recognition. They have also brought an algorithm for the selection of optimal images to generate eigenfaces.

In this approach face class is projected into a lower dimensional subspace where the mean square reconstruction errors for training images are minimum. The basis vectors for this subspace are called the principal components and in face analysis they take the name eigenfaces.

2.1.4.2.Neural Networks

Neural networks have widely been used in several pattern recognition problems. Face detection can be thought as a two class pattern recognition problem and neural networks have widely been used in this field. When trained with a large set of face images neural networks are shown to capture the complex class density of face class and perform quite well.

One drawback of neural networks is to choose optimum structure for neural network like number of hidden layers and their size. Care should also be taken in the selection of non face images (they should be close to face class) in order to improve the performance of training.

Propp and Sanal developed one of the earliest neural networks for face detection [9]. They used a 4 layer, 1024 input unit network for classification. H.A.Rowley, T.Kanade proposed a two stage neural network algorithm to detect faces in an image [5]. Their system is composed of two stage neural network filtering.

In the first stage a multiresolution search for face is done. Every point in the image is filtered using the first neural network. In order to detect faces with a larger size than search window, input image is subsampled and a search at this resolution is repeated using the same classifier.

A filter is applied to remove multiple and false face detections. Sliding window is used for counting positive results (extracted faces) in each region. Those regions with positive results greater than a threshold are classified as face and those not validating this condition are removed.

2.1.4.3.Support Vector Machines

Support vector machines (SVM) are shown to perform very effective face detection on cluttered scenes [12]. They perform structural risk minimization on training data to select the best decision boundary between classes. This decision boundary is indeed a set from the training data (support vectors) which are closest to the between class boundary. This fact is shown in Figure 2.3 where Support vectors are

shown in dark color. SVM require a huge amount of training data to select an affective decision boundary [12] and computational cost is very high even if we restrict ourselves to single pose (frontal) detection.

An early example to face detection with SVM was proposed by E.Osuna and R.Freund [31] in 1997. They worked with 19x19 face and non-face images. In their study SVM classification is combined with some priori image processing to increase the performance. They applied illumination gradient correction and histogram equalizations to get more silent images and decrease variations in training and test set face images. They reported correct detection rates of 97%.

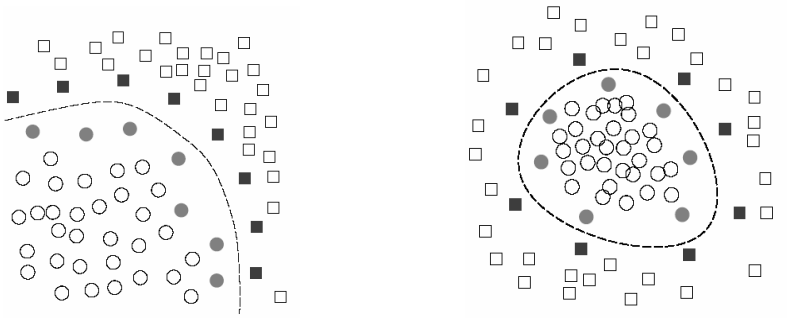


Figure 2-3 Support Vectors and Decision Boundary

2.2.Face Localization Stage of FAFE

The first stage of proposed facial feature extractor (FAFE) is to find location, size and orientation of face. Target images (or frames) are assumed to contain an over shoulder seen of a single person. With the help of this constraint the greatest ellipse on skin segmented binary image is selected as face. Skin segmentation and ellipse fitting will be described in the following sections.

2.2.1.Skin Segmentation

Human skin color has been used and proven to be an effective feature for face localization. Although different people have different skin color, several studies have shown that the major difference lies largely between their intensity rather than

their chrominance [7]. Several color spaces are proposed for color segmentation including RGB, normalized RGB, YCrCb, YIQ etc.

Normalized RGB which is also known as pure colors in the absence of brightness is the simplest color space which does not carry intensity but only color information. This color space is also reported to give quite accurate segmentation so it is selected in this study. Component b is redundant and does not provide any extra information since r+g+b=1. Pure colors are defined as follows:

$$r=R/(R+G+B) \quad (2.1)$$

$$g=G/(R+G+B) \quad (2.2)$$

$$b=B/(R+G+B) \quad (2.3)$$

Skin region samples from training set are collected and after being projected on (r,g) space, characteristics of their histogram are analyzed. Skin region distribution for used database is given in Figure 2-5. As can be seen from the figure, skin color distribution seem to stay in a narrow region on (r,g) space. This distribution shows that a simple level slicing on (r,g) values can result in an accurate segmentation.

Dark regions on image do not carry complete color information and cause false segmentations. In order to avoid this problem together with color thresholding a gray level thresholding is also applied on input images. In this study the color range used to extract skin region is taken as $0.4<r<0.5$ $0.2<g<0.4$ and the gray level threshold is taken to be $I>50$. Figure 2-4 shows the results of skin segmentation on some sample images.

Figure 2-4 Skin segmentation results using r,g & gray level threshold

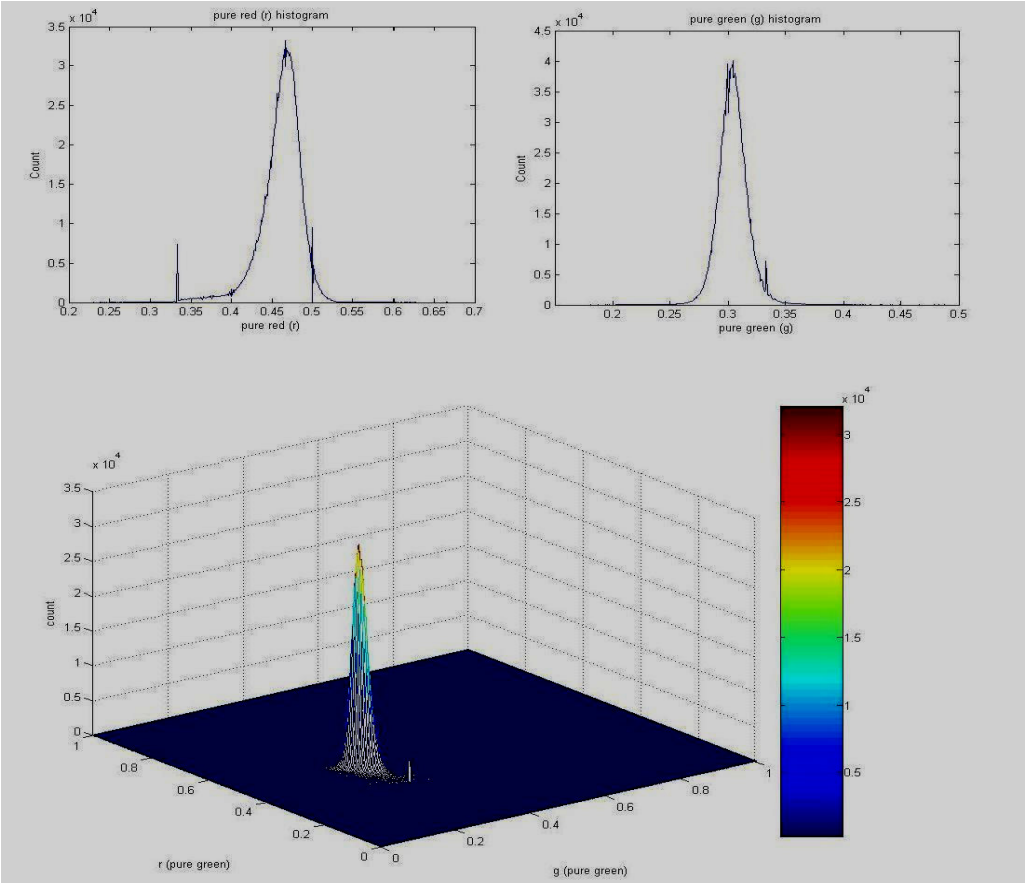


Figure 2-5 Histogram of Skin Color on (r,g) space

2.2.2.Ellipse Fitting

In this study input images are assumed to include frontal over shoulder view of a single person with allowable face rotation of 45 degrees. The method used to extract the angle, position and location of the face is to search for the largest possible ellipse in the skin extracted binary image. The ellipse region found is assumed as a valid face. Skin segmentation results in an output binary image including several small false detections and holes corresponding to facial organs, facial hair and glasses.

Morphological enhancements on skin extracted binary image are applied and false detections together with missing regions on face are corrected. Opening operation is

applied with a structuring element nxn where n equals to 1/32 of smaller image edge size. Information about morphological operations are included in Appendix-A.

Another problem with the binary image is the regions corresponding to facial organs and facial hair. Color distribution of those regions is clearly much different than skin color and they appear as holes on the extracted binary image. Closing operation with structuring element 3nx3n is applied on binary image to get a complete region for ellipse fitting. The next step in face localization is to search for a region that fits best ellipse. In order to make things easier, the largest rectangle which can fit in the extracted skin region is searched. Search for best fitting ellipse is later carried out on an extended version of this extracted region.

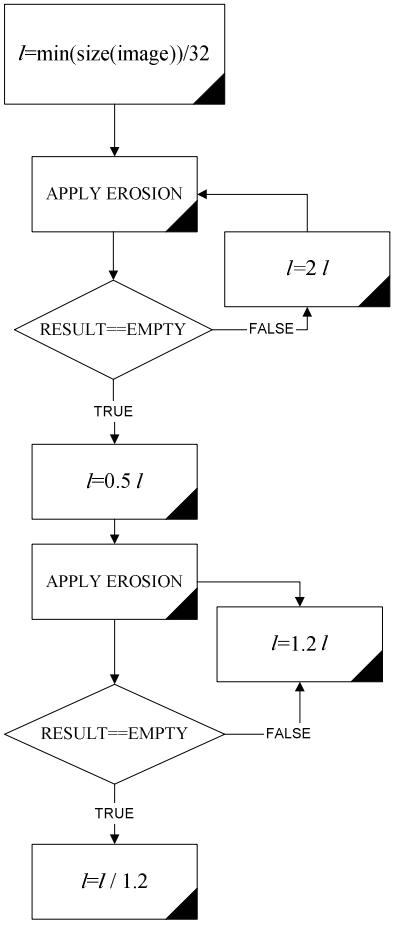


Figure 2-6 Algorithm to Find the Largest Rectangle on Skin Segmented Images

Search for the largest rectangle is done by applying erosion operation on binary image with a rectangle of size 3*l*x2*l* where the value of *l* is iteratively increased using an algorithm shown on Figure 2.6.

E.Saber and M.Tekalp [6] proposed a method to extract facial features using color shape and symmetry based cost functions. They modeled face as an ellipse and found most representative ellipse using the variations in vertical and horizontal directions. Face ellipse fitting technique used in this thesis is essentially the same as the technique proposed in [6].

After finding the region to search for an ellipse the following algorithm is applied to extract face ellipse parameters for the face region. A sample rotated ellipse like boundary, parameters used for ellipse fitting and optimal ellipse is shown in Figure 2.7.

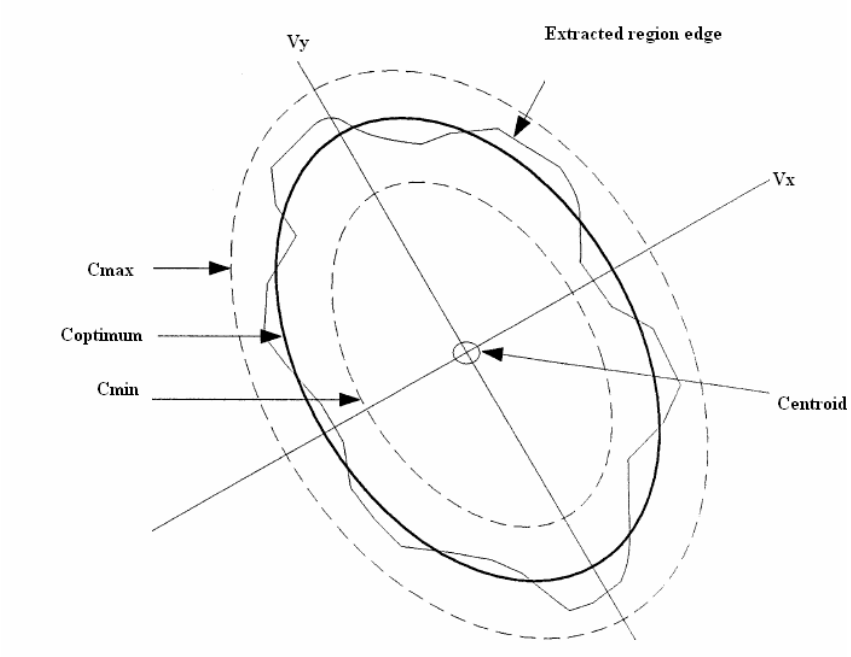


Figure 2-7 A Sample Ellipse Like Boundary and Best Fitting Ellipse

Assume the center of mass for the ellipse like shape given in Figure 2.7 is represented with (C_x, C_y). The variation in x directions σ_x^2 and the variation in y direction σ_y^2 can be calculated using

$$\sigma_x^2 = \sum_i^D (X_i - C_x)^2 \Big/ D^2 \tag{2.4}$$

$$\sigma_y^2 = \sum_i^D (Y_i - C_y)^2 \Big/ D^2 \tag{2.5}$$

Similarly the cross variance σ_{xy} can be calculated using

$$\sigma_{xy} = \sum_i^D (Y_i - C_y)(X_i - C_x) \Big/ D^2 \tag{2.6}$$

By combining these terms covariance matrix of target region can be written as

$$R = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \tag{2.7}$$

"The eigenvalues (λ_x, λ_y) of R provide us with a reasonable estimate of the spread of the skin region in the directions of the corresponding eigenvectors (V_x, V_y)" [6]. The directions of the eigenvectors indicate the principal axis of the ellipse candidate region. If $V_x = \begin{bmatrix} Vxa \\ Vxb \end{bmatrix}$ orientation of the face is given by the following equation:

$$\theta = \arctan(Vxa/Vxb) \tag{2.8}$$

Next step in ellipse fitting algorithm is to rotate extracted skin region with angle θ. Resultant image is edge filtered and best ellipse of the form:

$$\frac{(x - C_x)^2}{\lambda_x^2} - \frac{(y - C_y)^2}{\lambda_y^2} = c \tag{2.9}$$

is found.

In order to find this ellipse first a range of values for constant c is required. A search range [Cmin Cmax] for the best value of c is defined using the boundary of skin region. The logical choice for Cmin is the smallest ellipse which passes from the nearest point to center point of skin region. Similarly Cmax is chosen using the most distant point to center of skin image.

Mathematically these values can be found as follows. If the function $f(x, y)$ defined as

$$f(x, y) = \frac{(x - C_x)^2}{\lambda_x^2} - \frac{(y - C_y)^2}{\lambda_y^2} \tag{2.10}$$

is applied on the edge of searched region, the largest and smallest values of $f(x, y)$ gives us Cmin Cmax values. Cmin Cmax values for an example boundary are also shown in figure 2-7.

In the defined range of parameter C, we compare the modified Hausdorff distance between extracted skin region edge and active ellipse boundary for several values of C. The value of C which gives the minimum modified Hausdorff distance will be chosen as the best ellipse for the face region.

Hausdorff distance can be defined as follows:

Given two set of points S1={a_0 , a_1 ,…, a_n } and S1={b_0 , b_1 ,…, b_n } Hausdorff distance

$$H(S1,S2) = \max(h(S1,S2) ,h(S2,S1) \tag{2.11}$$

$$h(S1,S2)= \max_{i \in S1}(\min_{j \in S2}(\|i - j\|)) \tag{2.12}$$

A modified version of Hausdorff distance can be defined as

$$h(S1,S2)= \underset{i \in S1}{mean}(\min_{j \in S2}(\|i - j\|)) \tag{2.13}$$

Hausdorff distance is a direct measure of greatest distance between two point set. Skin extraction may result in deviations from actual face ellipse caused by false extraction and regions like ears and neck. Modified Hausdorff distance averages distance over all point set improving performance on such images. Because of this property modified Hausdorff distance is preferred to standard Hausdorff distance in this study.

After finding the best ellipse for the skin region, angle of the face is corrected and together with face size and location this corrected image is used in feature extractor

subsystem. A block diagram of face locator subsystem is given in the Figure 2-8. A sample image and results of the algorithm on this image is given in Figure 2-9.

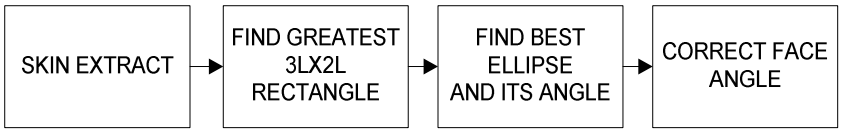


Figure 2-8 Face localization Block Diagram

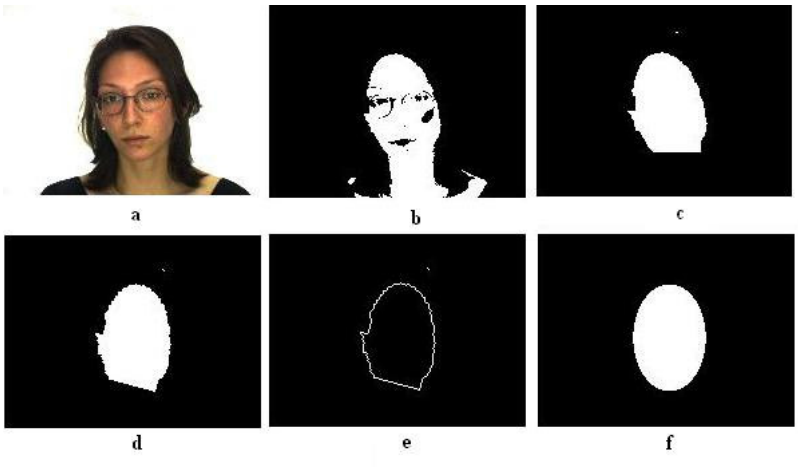


Figure 2-9 Face Localization Results on Sample Image

First skin extraction is applied on input image (Figure a). Resultant image (Figure b), includes holes corresponding to facial organs and hair. Holes on the image are closed using a dilation followed by an erosion operation (Figure c). Angle of the image is found and corrected to 90 degrees (Figure d). Edge of the angle corrected image is found and the best ellipse is searched on the range [Cmin Cmax] which was found as described above.

In defined range of ellipse sizes modified Hausdorff distance between the edge and ellipse boundaries are compared and the ellipse with the minimum Hausdorff distance is assumed to be the best ellipse describing the face region. Change of

Hausdorff distances for sample image in Figure 2.9 and changing c values is given in the graph below.

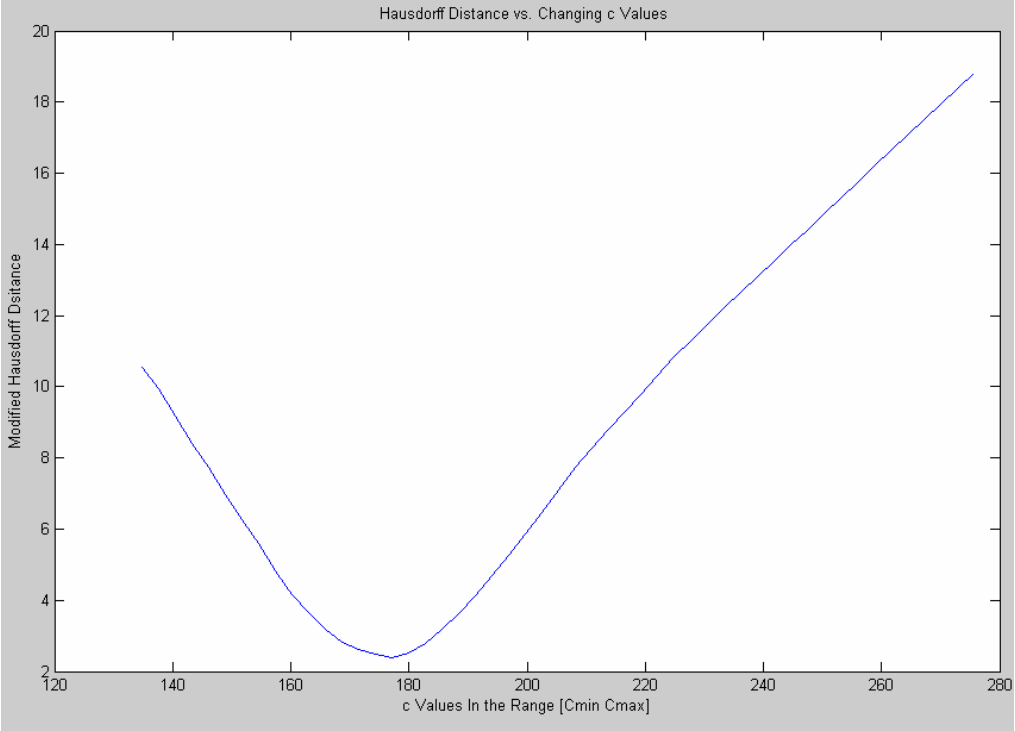


Figure 2-10 Hausdorff Distance for changing c in the [Cmin Cmax] Range

For the sample image given in Figure 2-9-a Hausdorff distance between actual skin region and candidate ellipse is analyzed for all possible values of c in the range [Cmin Cmax]. C value giving the minimum Hausdorff distance (in this case 177) is assumed to define the best ellipse fitting to the extracted skin region. Extracted ellipse is shown in Figure 2-9-f.

CHAPTER 3

FACIAL FEATURE EXTRACTION

Facial analysis studies can be separated into two main classes which are holistic and feature based methods. Holistic methods analyze faces as a whole without partitioning it into regions. Feature based techniques locate features belonging to facial organ boundaries or calculate parameters defining their shape.

The feature extractor proposed in this study (FAFE) can be thought as a two step algorithm. First step calculates rough location of COE and COM in low resolution images. This information is used to define search windows for upper and lower facial regions. Feature coordinates are extracted using a template matching approach where template parameters are calculated using PCA.

First, some examples to available methods will be described in section 3.1. Section 3.2 describes 3 different techniques to extract center of eyes and a single technique to locate mouth. Section 3.3 describes upper and lower facial region localization and template matching process. A facial feature tracking technique based on KLT tracker which uses FAFE for initialization and correcting lost feature points will be described in section 3.4. Expression classification using extracted features is described in Section 3.5.

3.1.Facial Feature Extraction Techniques

Boundaries of facial organs like eyes, eyebrows, mouth are most widely used facial features in face analysis. Face recognition, lip tracking, eye tracking can be given as examples to popular subjects which directly aim to follow changes on facial boundaries.

Facial feature is most commonly used as a generalized name to facial boundaries and these two names will be used interchangeably in this text.

Most common techniques to extract facial features can be classified as follows:
1.Template Matching
2.Wavelet Based Techniques
3.Region Segmentation Based Techniques

Among the three group of techniques template matching is perhaps the most widely used one. Template matching as it was described in section 2.1.3.2 is based on deforming a representative template by minimizing a cost function which is a measure of dissimilarity between current template shape and target region.

A predefined shape for searched object is an important information for template matching, however facial boundaries show wide range of variations among different people. Expression, gender, occlusion by facial hair or glasses also add variations to boundaries on face image. With these difficulties defining a predefined shape for templates is a challenging task.

M.Malciu [36] used a simplified version of MPEG-4 face model where mouth is modeled using 4 lines describing boundaries of upper and lower lips. Eyes are modeled using 4 points.

Y.Tian and Takeo Kanade [32][35] proposed a multi state template matching algorithm for extracting facial features. In this study three templates representing mouth in three different shapes corresponding to open, closed and tightly closed mouth are used. They initialized template on the first frame and updated template parameters in following frames. Active mouth template in current frame was selected using color information. Similarly they modeled eyes with a two state template corresponding to open and closed eyes. Active template model was selected using an iris detector.

A.Kapoor and W.Picard [1] combined infrared based eye detection and template matching for extraction and tracking of upper facial features. They used PCA for recovering template parameters and report a feature tracking rate of 30 fps. Feature extraction technique implemented in this system is an extension of this technique.

Most of the template matching approaches require an initialization of template parameters in the first frame of image sequence. Studies described in [1] do not require initialization, however proposed system uses an infrared camera to detect eyes. After locating eyes, template describing upper facial features is located on upper face and template parameters are calculated.

In this study an alternative method for locating the upper face template is proposed. Instead of detecting eyes, COE is selected as a feature to locate upper face template.

After finding face location together with its size and orientation, key facial coordinates corresponding to upper facial features from eyes, eye brows and lower facial features from mouth region are extracted. Two separate templates for UFR (Upper Facial Region) and LFR (Lower Facial Region) are used for extracting feature locations in these regions. Template parameters are calculated using PCA once they are placed to COE and COM.

3.2.Center of Eyes (COE) and Center of Mouth (COM) Extraction

Feature extraction starts with locating the COE on the image. Three different algorithms to find the COE are implemented and their performances are evaluated. These three algorithms can be listed as follows:

1.Using a special filtering technique which takes advantage of two dark regions next to the region between eyes.

2.Using a region based technique which takes advantage of symmetry of the face

3.PCA (Principal Component Analysis) search.

First two algorithms use features around COE and they can be thought as feature based methods. PCA search analyzes eyes and eyebrows as a whole and tries to find the best match for eyes region.

One common and fast way of locating facial features is to search for a reference point on the face and defining search window for each facial feature. Feature point search can be carried out in a faster and more reliable manner once search windows are defined. Center of eyes (COE) point which can be defined as the center of 4 upper facial regions (eyes and eyebrows) is a good choice for such a reference point. COE is chosen as a reference point due to the following properties it carries:

1.It can be seen on a wide range of face orientations.

2.It does not change in different expressions on the face.

3.Occlusion by facial hair, make up or glasses is not common.

Three different methods to extract COE are analyzed. These methods are described in the following section.

3.2.1.Feature Based Methods to Extract Center of Eyes

Shinjiro Kawato, Jun Ohya [3] proposed a method to extract COE with a new filtering technique named as circle frequency filter. If we investigate the gray level distribution of center of eyes neighborhood we find two bright regions corresponding to upper and lower skin regions and two dark regions corresponding to eyes and eyebrows. Circle frequency filter takes advantage of this property. The technique can be described as follows.

Let $f_r(x, y) = \{i_1, i_2, ..., i_n\}$ be the gray level intensities of points on the circle with center (x, y) and radius r where i_1 corresponds to the point just above the circle center. If $f_{(x,y,r)k}$ is the intensity of the k'th point on the circle with center (x, y) and radius r, discrete Fourier transform (DFT) of these points and circle filter output $CF_r(x, y)$ is given as follows:

$$CF_r(x, y) = \sum_{k=0}^{N-1} f_{(x,y,r)k} e^{2\pi ikn/N} \tag{3.1}$$

Circle frequency filter is basically the second frequency component of DFT which gives peak on circle centers with gray level values corresponding to two cycles of a sinusoid.

The data set for correct center of eyes point starts with a high value and a low high low intensity pattern follows. Magnitude of circle filter gives us the probability of a two cycle sinusoid in data set but does not give any information about its phase. Those points which have a similar intensity pattern with COE will have $\mathrm{Re}(CF(x,y)) > 0$.

Since skin region between eyes and eyebrows form a bright region between two dark regions a median filtering is applied on images to form two continuous dark regions on both sides of center of eye region. Circle for correct center of eye is shown in Figure 3.2 and gray scale values on this circle are given in Figure 3.1.

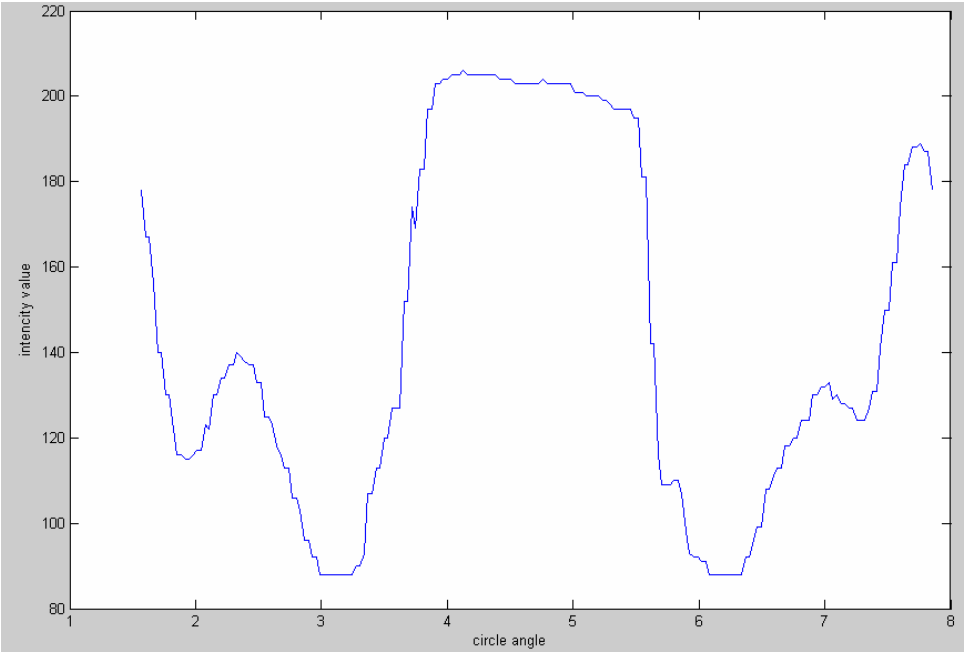


Figure 3-1 Intensity Pattern of COE Circle on a Sample Image

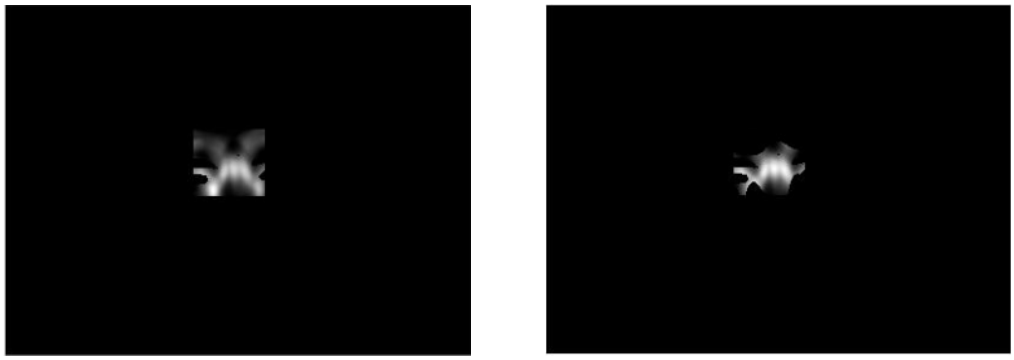
Figure 3-2 Median Filtered Sample Image for COE Extraction

By using the parameters of the extracted face ellipse, we define a region to search for center of eye region. Figure 3-3-a gives the magnitude of circle filtering on the search region for sample image given in Figure 3-2.

In order to select regions with similar intensity pattern with COE, the filter given by

$$H_{PHASE\_MATCH}(x,y) = \begin{Bmatrix} 1 & if & \mathrm{Re}(\mathrm{CF}(x,y)) > 0 \\ 0 & & else \end{Bmatrix} \tag{3.2}$$

is applied to circle filtered image and the result is shown in Figure 3-3-b.



a b

Figure 3-3 Circle Filtered and Phase Filtered Images.

The maximum of result is assumed to be the center of eye. Figure 3.2 gives the extracted center of eye point on a sample image.

E.Saber and M.Tekalp [6] proposed to use symmetry-based cost functions to locate eyes, nose and mouth within the facial segmentation mask. They defined non skin regions on face region as ‘‘holes’’. The aim of this technique is to find the two ‘‘holes’’ within the face mask that are most likely to be the eyes, and to locate the nose and mouth once the eyes have been identified.

In this study an extended version of this method is implemented and its performance is evaluated. Although not included in [6] eyebrows also appear to be extracted as holes on skin extracted images. Cost functions used in [6] are modified to count for eyebrows and used in this study.

Face exhibits a high amount of symmetry which can be used to discriminate between face and non-face regions on images. This symmetry is also a very important clue for the position of facial organs which are located symmetrically around major and minor axis of face. Symmetry properties and corresponding cost functions are given below.

Let (C_x, C_y) be the extracted ellipse center, b be the ellipse major axis length and $(H1_x, H1_y)$ and $(H2_x, H2_y)$ be the centers of the two holes.

1.Line connecting the eye centers is parallel to face ellipse minor axis. As a result eyes have similar distance to minor axis.

$$Cost_i(H_1, H_2) = \left\| H_{1y} - C_y \right| - \left| H_{2y} - C_y \right\| \tag{3.3}$$

2.Eyes are symmetric with respect to the major axis of face ellipse so has a similar distance to it:

$$Cost_{ii}(H_1, H_2) = \left\| H_{1x} - C_x \right| - \left| H_{2x} - C_x \right\| \tag{3.4}$$

3.Eyes are the closest holes in the face to the minor axis and their total distance to major axis can be used as a part of cost function.

$$Cost_{iii}(H_1, H_2) = \left| H_{1x} - C_x \right| + \left| H_{2x} - C_x \right| \tag{3.5}$$

4.Eyes are at different sides of major axis

$$Cost_{iV}(H_1, H_2) = \left| sign(H_{1x} - C_x) - sign(H_{2x} - C_x) \right| \quad (3.6)$$

5.Eye brows are closer to the upper facial ellipse boundary than eyes.

$$Cost_V(H_1, H_2) = \left| H_{1y} - b/2 - C_y \right| + \left| H_{2y} - b/2 - C_y \right| \quad (3.7)$$

6.Every possible pair is analyzed and those conflicting with $Cost_{iV}(H_1, H_2)$ =0 are removed. Remaining pairs are assigned a weighted cost function of the form

$$Cost_{eyes}(H_1, H_2) = Cost_i(H_1, H_2) + 5\,Cost_{ii}(H_1, H_2) + Cost_{iii}(H_1, H_2) \quad (3.8)$$

The pair with smallest $Cost_{eyes}(H_1, H_2)$ is chosen as the eye pair. Eye pair is removed from the holes list and remaining hole pairs are assigned the following cost function

$$Cost_{eyesbrows}(H_1, H_2) = Cost_i(H_1, H_2) + Cost_{ii}(H_1, H_2) + Cost_v(H_1, H_2) \quad (3.9)$$

The pair minimizing this function is chosen to be the eyebrow pair.

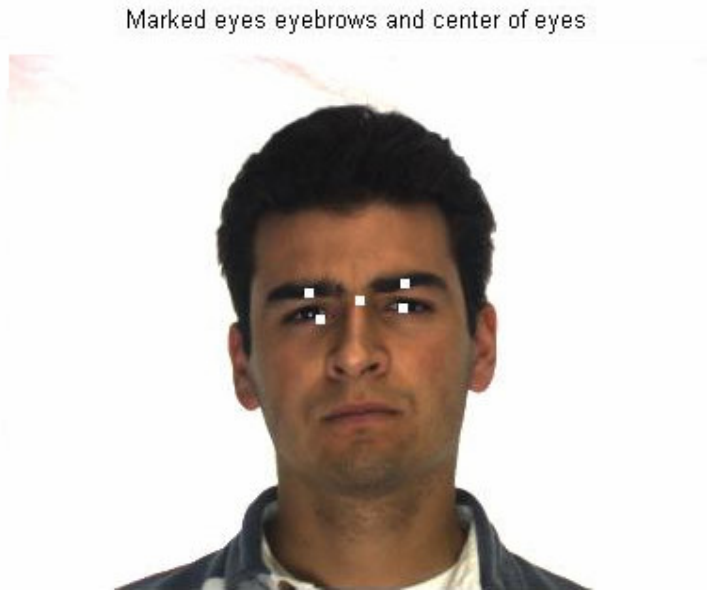


Figure 3-4 Extracted Facial Coordinates Using Region Based Technique

Although this method seems to be easy it is not robust to partial occlusion caused by glasses and facial hair. Also eye regions are hard to be found on low resolution images when eyes are closed. Extracted facial regions using cost function based algorithm is shown in Figure 3-4.

3.2.2.COE Localization Using PCA

Another way of center of eyes detection is PCA. General Information about PCA is presented in Appendix B. First usage of PCA for detecting members of a class will be described. Implementation of this technique for UFR and LFR localization is presented in following sections.

Principal component analysis can be used for detecting the members of a class in an image. Assume training samples from a class Γ is collected and principal components $U = [u_1 u_2 \quad . \quad . \quad . \quad u_k]$ of this sample set are calculated. Weighted sum of these principal components span a subspace.

If the number of training samples and used principal components are large enough, every member Γ_x of class Γ can be approximately described as a weighted sum of u_i where weights are projection of Γ_x (with coordinates given by $\alpha = [\alpha_1 \alpha_2 \quad . \quad . \quad \alpha_k]$) on constructed subspace. If we reconstruct Γ_x using

$$\tilde{\Gamma}_x = \sum_{j=1}^{K} (\alpha_j u_j) + \mu \tag{3.10}$$

and define reconstruction error ε as

$$\varepsilon = \left\| \Gamma_x - \tilde{\Gamma}_x \right\|^2 \tag{3.11}$$

ε will be low when the number of principal components and training samples are sufficient. Projection of a sample which is not a member of Γ would result in a large reconstruction error. This means ε can be thought as a measure of distance to the searched class. If we calculate reconstruction error on the whole image, those regions corresponding to low ε can be selected as regions which are close to searched class.

An idea similar to this approach with its application to eye region extraction and tracking was used in [4].

The methods described in 3.1.1 require less calculations and circle filtering technique has shown to be a good choice for a real time system [5]. But they are highly sensitive to partial occlusion of face. Facial hair and glasses prevent correct region segmentation of face and false regions can be extracted as eyes using symmetry based cost functions. Similarly shadows and hair partially closing face can generate regions having similar neighborhood with COE preventing circle filtering function to give maximum at correct location.

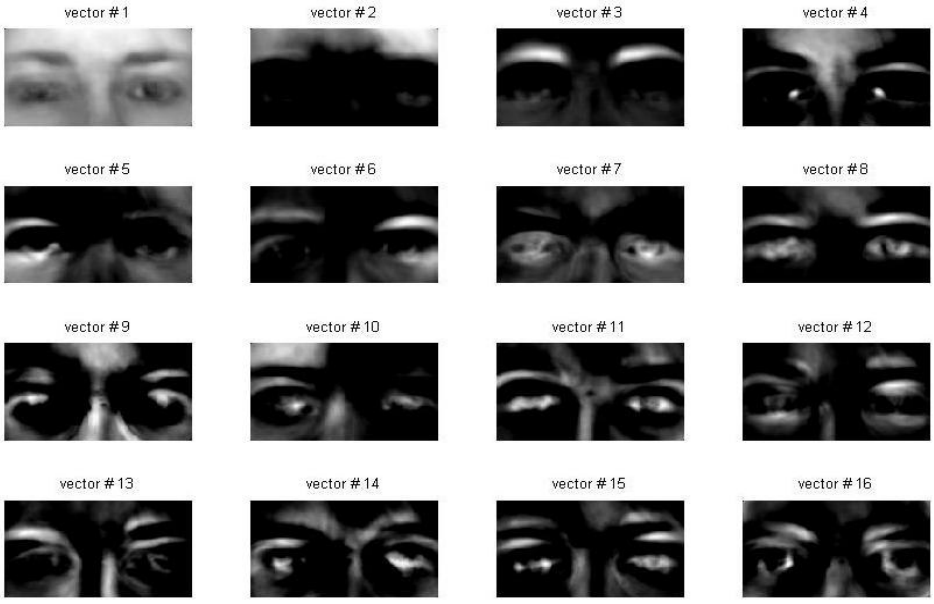
In this study PCA is used to model upper facial region (UFR) which can be thought as a pattern class on its own. Localization of UFR is done by projecting candidate face regions into principal components subspace and searching for the region with minimum reconstruction error as described in previous section. The center point of found region is used as the COE.

The number of training images and the number of principal components are important parameters affecting the coverage performance of used principal components in analyzed subspace. Similarly illumination variations and noise in analyzed images also degrade performance of PCA by increasing within class scatter of UFR.

Another parameter effecting coverage performance is the scale variations in training and test sets. Multi scale UFR patterns can be localized using principal components which were calculated using training patterns at various scales. However such an analysis would require significantly greater number of principal components which cause an increase in the number of required training patterns and computation time. In order to avoid this shortcoming training patterns are first brought to same scale and principal components are calculated using these scaled patterns and technique described in Appendix-B. For bringing each training pattern to the same scale short axis length of extracted face ellipses are brought to 230 pixels. Sixteen eigen images corresponding to the largest eigenvalues are selected

for usage. Figure 3-5 shows calculated mean image (b) and 16 principal components used to search COE (a).

Searching the whole face ellipse for localizing best UFR is a time consuming operation. Computation time can be decreased using a multiscale search on face region. Approximate localization can be done using low scale principal components and fine tuning can be done by using higher scale ones.



a

b

Figure 3-5 Eigen and mean Images for COE detection

COE localization steps as they are implemented in this study are given in Figure 3-6. Input image is first resized such that the width of the resultant image is 80 pixels. Skin extraction and ellipse fitting operations are performed on this image (B).

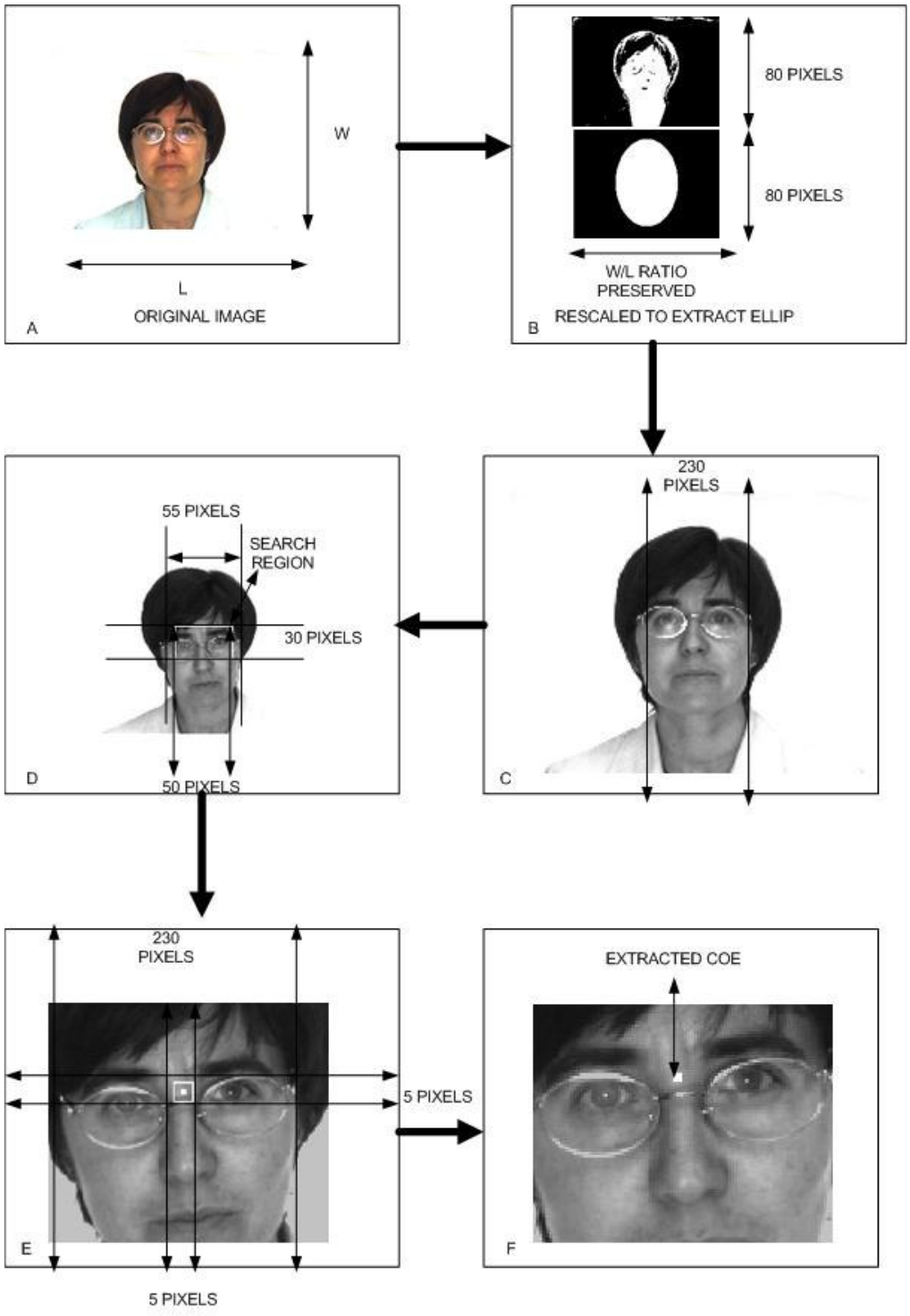


Figure 3-6 COE Extraction Steps

After ellipse fitting, the localization of face and its approximate size is known. This information is used to resize face image such that it has a face ellipse with short axis length equal to 230 pixels (C). By this operation scale variation between training patterns and target face image are removed. As a direct result principal component also has the same scale with target image.

Although face ellipse gives us highly accurate information about face's horizontal length and position vertical coordinate of face center and extend in vertical direction is usually not accurate. This fact is caused by possible partial occlusion of face region from top or bottom. An example to this situation can be seen from the sample image in Figure 3-6 (A) where hair occludes one third of the face region.

To solve this inaccuracy, UFR search on face region is performed on a relatively large region. The computation time for UFR localization is decreased using a two step search on face. In the first step face image and principal components are resized by ¼, rough location of UFR is calculated using these components.

The region of search chosen is a rectangle defined by its upper left (X_{ULE}, Y_{ULE}) and lower right (X_{LRE}, Y_{LRE}) corners. If (Xc, Yc) is the center of face ellipse then rectangle corners are given by

$$X_{ULE} = X_C - 25 \tag{3.12}$$

$$X_{LRE} = X_C + 25 \tag{3.13}$$

$$Y_{ULE} = Y_C - 30 \tag{3.14}$$

$$Y_{LRE} = Y_C \tag{3.15}$$

The point corresponding to lowest reconstruction error is later used for defining a second search window of 5x5 pixels on full scale image (E). Original principal components are used to locate the UFR with higher accuracy and the center of localized UFR is extracted as COE.

3.2.3.COM Localization Using PCA

After extracting the COE location, a rough estimate for mouth region can be found with the help of extracted ellipse parameters. The region of search for mouth is given in the Figure 3-7.

Figure 3-7 Center of Mouth (COM) Search Region On a Sample Image

If the COE location is given by $(Xcoe, Ycoe)$ then the chosen mouth search region upper left (X_{ULM}, Y_{ULM}) and lower right corners (X_{LRM}, Y_{LRM}) are given a follows:

$$X_{ULM} = Xcoe - 15 \tag{3.16}$$

$$X_{LRM} = Xcoe + 15 \tag{3.17}$$

$$Y_{ULM} = Ycoe - 10 \tag{3.18}$$

$$Y_{ULM} = Ycoe - 50 \tag{3.19}$$

Similar to the center of eye localization mouth localization is also done using PCA. Calculated mean image and Eigen images for mouth region are given in the Figure 3-8 and 3-9 respectively.

Reconstruction error ε is calculated on the search region and the point corresponding to minimum ε is chosen to be the best candidate for mouth location. Extracted location for mouth region on a sample image is given in Figure 3-10.

Figure 3-8 Calculated Mean Lower Facial Region

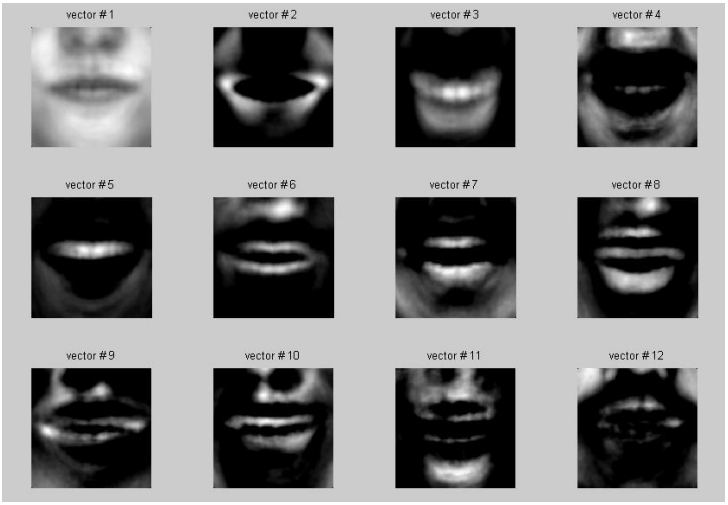


Figure 3-9 Principal Components For Lower Facial Region

Figure 3-10 Extracted Center of Mouth (COM)

3.3.Facial Feature Extraction Using PCA Based Template Matching

Principal component analysis not only gives us the probability of a class membership but also gives us the ratios of variations from the mean image in the directions of principal components. One application of this property is template parameter recovering using the coordinates of the analyzed image on principal component subspace. ($\alpha = [\alpha_1 \; \alpha_2 \quad . \quad . \quad \alpha_k]$)

A.Kapoor and W.Picard [1] used an upper face template constructed using 22 facial features. Template parameters are recovered using PCA analysis. In this thesis, a simplified version of their template is used for upper facial feature extraction. Template parameter recovering technique is identical with the one described in [1]. Their technique is extended for a mouth template and a similar analysis is carried out to calculate its parameters.

In this study a face template representing eyes, eyebrows and mouth boundary is assumed to be a sufficiently descriptive model for facial expression analysis. This template is given in the Figure 3-11.

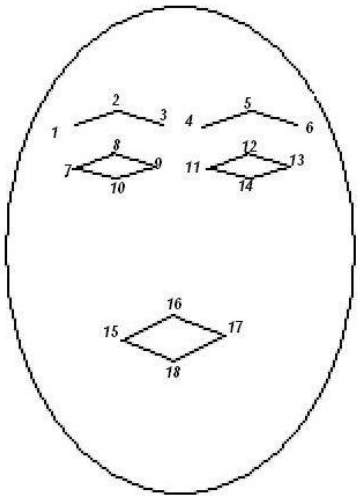


Figure 3-11 Used Face Template

Face template is constructed using separate eye, eyebrow and mouth templates constructed from 18 facial feature points coordinates (14 upper facial features, 4 lower facial features) which are described below:

Eye brow template: Leftmost, Rightmost and Top coordinates of eyebrow region
Eye Template : Leftmost, Rightmost, Top and Lowest coordinates of eye region.
Mouth Template: Leftmost, Rightmost, Top and Lowest coordinates of mouth region

From the basics of PCA, we know the fact that if an image class is represented with sufficient number of principal components, every member of this class can be approximately described as a linear combination of principal components plus class mean. The weights used in this linear combination are the coordinates of the image in the principal subspace.

If this class of images contains a set of feature points of interest then every principal component will also carry variance information for those feature points. This information will be called as coordinate variance (CV). Similarly mean image of the class will carry the mean coordinate information for those feature points. This value will be called as mean coordinate.

If we analyze equation B.14 (In Appendix-B), we can see that every eigen image is also a weighted sum of training set images after removing the mean image from each one. So CV of each eigen image can be calculated using the same idea.

The technique to calculate CV in this study is using the same weights in 3.23 after removing mean coordinate from feature coordinates of the training set. The following procedure is applied to extract upper and lower facial feature coordinates
14 upper and 4 lower facial features are hand marked on training set images
14 coordinates for upper facial region of image I_i are recorded in array

$$C_{li} = [p_1 \quad p_2 \quad . \quad p_{14}]$$

4 coordinates for lower facial region of image I_i are recorded in array

$$C_{ui} = \begin{bmatrix} p_{15} & p_{16} & p_{17} & p_{18} \end{bmatrix}$$

Mean of the coordinates is calculated using

$$C_{um} = \frac{1}{M}\sum_{j=1}^{M} C_{uj} \tag{3.20}$$

$$C_{lm} = \frac{1}{M}\sum_{j=1}^{M} C_{lj} \tag{3.21}$$

Array of mean removed feature coordinates is constructed:

$$C_L = \begin{bmatrix} C_{l1} - C_{lm} & C_{l2} - C_{lm} & . & C_{lM} - C_{lm} \end{bmatrix} \tag{3.22}$$

$$C_U = \begin{bmatrix} C_{u1} - C_{um} & C_{u2} - C_{um} & . & C_{uM} - C_{um} \end{bmatrix} \tag{3.23}$$

Similar to equation B.14 (Appendix-B) CV for each eigen image Pi is calculated using:

$$\mathrm{P}_{\mathrm{iL}} = C_L w_i \tag{3.24}$$

$$\mathrm{P}_{\mathrm{iU}} = C_U w_i \tag{3.25}$$

where w_i is the i'th eigen vector of covariance matrix used for principal component calculation.

After projecting extracted upper facial region to principal component subspace (calculated for upper facial region) coordinates on this subspace are used to extract upper facial feature coordinates on image $\mathrm{I_i}$ using:

$$\tilde{C}_{Ui} = \sum_{j=1}^{K} (\alpha_j \mathrm{P}_{\mathrm{iU}}) + C_{um} \tag{3.26}$$

Similarly after projecting extracted lower facial region to principal component subspace (calculated for lower facial region) coordinates on this subspace are used to extract lower facial coordinates on image $\mathrm{I_i}$ using:

$$\tilde{C}_{li} = \sum_{j=1}^{K} (\alpha_j \mathrm{P}_{\mathrm{iL}}) + C_{lm} \tag{3.27}$$

Facial feature extraction result for a sample image from AF database is given in Figure 3-12.

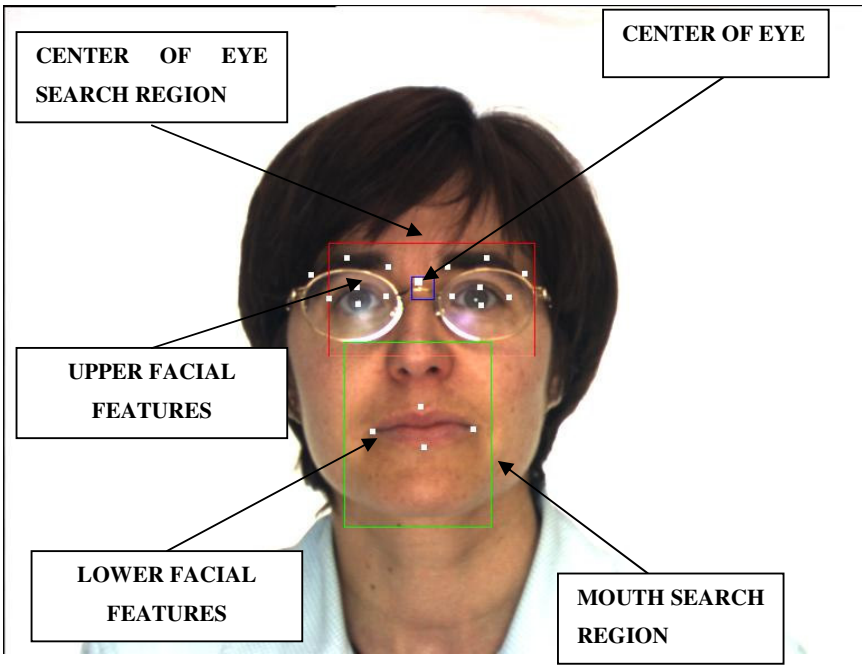


Figure 3-12 FAFE Results On a Sample Image

3.4.Facial Feature Tracking Using KLT Tracker and FAFE

Similar to extraction of facial features, tracking of extracted feature regions is also an important problem for facial expression analysis. Feature extraction processes are usually more time consuming then tracking extracted features, so a successful feature tracker can drastically improve timing of expression analysis algorithms. Similarly temporal location of feature points together with order of their movement can be used as powerful clues for expression recognition.

It is possible to make generalizations for common expressions like surprise smile, fear which define feature point changes (rising eyebrow, extending of mouth in horizontal direction, closed eyes) for those expressions. Occurrences of such changes can be detected using thresholds for each predefined movement.

Several techniques for facial feature tracking have been proposed in literature. However most of these techniques require a manual initialization of tracked feature points. An important problem in feature tracking is the possibility of losing target

features. Feature points can significantly change shape or they may get temporally occluded during tracking, avoiding tracker to follow them. A technique to relocate lost feature points is also an important requirement for facial feature tracking.

Although analysis of facial expression changes in video sequences is not a subject of this study, proposed feature extractor FAFE can be used as an initialization and relocating technique for facial feature locations.

A facial feature tracking system using KLT tracker and FAFE feature extractor is proposed and its performance on sample video sequences are evaluated. Extraction of facial expressions on video sequences is not included in this study.

Section 3.4.1 will briefly describe KLT tracker. Section 3.4.2 describes criteria for choosing good features to track with KLT tracker. An evaluation of face images using the described criteria is also included in this chapter. Section 3.4.3 will describe proposed feature extractor together with rules to detect lost features.

3.4.1.Kanade Lucas Tomasi (KLT) Tracker

Assume a video sequence containing a moving object in frames. Every frame of this sequence includes a region (corresponding to the moving object) which can be expressed as a translated version of the same region in previous frame.

Let t be the frame index and x, y correspond to vertical and horizontal coordinates. If only a displacement is assumed and moving object does not rotate the following relation hold for frames I(x, y, t).

$$I(x, y, t) = I(x + d_x, y + d_y, t + \tau) + e(x, y) \tag{3.28}$$

The $e(x, y)$ term rises from several facts. The boundaries of interested object can appear and disappear in frames due to occlusion by other objects. Similarly illumination conditions for two frames can significantly differ causing an intensity difference between regions in different frames.

With the effect of error term, tracking of a single point in images is usually impossible if it does not have a significantly different or unique chrominance. Tracking regions defined by some window is a solution to this problem however points belonging to a window may behave differently. Different points belonging to the same window can move in different speeds or they may disappear and reappear independently.

Window size is a parameter limiting variations in the window by changing region size. If we choose a relatively small window, behavior of included points would be similar and an assumption of constant displacement can be made on this window.

Let $\bar{x} = (x, y)$ define the horizontal and vertical coordinates and $\bar{d} = (d_x, d_y)$ be the displacement between two frames (I and J). If there is no rotation but only a displacement equation 3.28 can be rewritten in vector form as $\mathrm{I}(\bar{x}) = \mathrm{J}(\bar{x} + \bar{d}) + e(\bar{x})$
A simple measure of error in displacement calculation is as follows:

$$\varepsilon_{\bar{d}} = \int_{w} \left[\mathrm{I}(\bar{x} - \bar{d}) - \mathrm{J}(\bar{x})\right]^2 w(\bar{x}) d\bar{x} \qquad (3.29)$$

$w(\bar{x})$ is a weighting function which defines the importance of each point for error calculation. A common choice for $w(\bar{x})$ is a Gaussian function which emphasizes the importance of window center [33].

A straightforward technique to find optimal value of displacement vector is to search the whole possible values of $\bar{d}$. This technique is very time consuming and calculation time for an (MxM) window and (NxN) possible displacement space grows with $M^2 N^2$.

Hill climbing technique is an alternative to straightforward search. In this technique and initial guess for displacement vector is made and error function for surrounding displacements are calculated, next guess for displacement vector is done in the

direction of minimum error. This procedure is repeated until a convergence measure is met.

Kanade- Lucas-Tomasi (KLT) used a slightly different interpretation of equation 3.29 for defining the cost function and calculated the displacement vector by minimizing this cost function. The solution of Lucas Tomasi Tracking equation is using Taylor series expansion is given by the expressions stated below:

$$\varepsilon_{\bar{d}} = \int_{w} \left[\mathrm{I}(\bar{x} + \frac{\bar{d}}{2}) - \mathrm{J}(\bar{x} - \frac{\bar{d}}{2}) \right]^2 w(\bar{x}) d\bar{x} \tag{3.30}$$

$$g = \begin{bmatrix} \frac{\partial}{\partial x}(I + J) \\ \frac{\partial}{\partial y}(I + J) \end{bmatrix} \tag{3.31}$$

$$Z = \int_{w} \mathrm{g}(\bar{x}) \mathrm{g}(\bar{x})^{\mathrm{T}} w(\bar{x}) d\bar{x} \tag{3.32}$$

$$e = 2 \int_{w} [\mathrm{I}(\mathrm{x}) - \mathrm{J}(\mathrm{x})] \mathrm{g}(\bar{x}) w(\bar{x}) d\bar{x} \tag{3.33}$$

$$Z\bar{d} = e \tag{3.34}$$

3.4.2.Feature Selection for KLT Tracking

Texture is an important parameter which changes the availability of motion information in video frames. A continuous horizontal edge is a suitable texture for calculating vertical motion, however horizontal motion of this edge can not be calculated since it does not cause any change in window.

Calculation of reliable motion from images can be guaranteed by selecting rich enough textures. Several measures to select textures for tracking were proposed. Selecting corners, selecting regions with high spatial frequency content or high second order derivatives are some examples to used measures [34].

T. Kanade and C. Tomasi proposed a different measure optimal for KLT tracker. For the reliable solution of Equation 3.34, “the 2x2 matrix Z must be above image

noise level and it must also be well conditioned. Noise criteria correspond to large eigenvalues requirement for Z. Well conditioning can be met by selecting regions with similar scale eigenvalues" [34].

The relation between eigenvalues of Z and texture can be stated as follows. If both eigenvalues are small then corresponding texture has approximately continuous intensity and extraction of reliable horizontal and vertical motion is not possible. If there is a significant scale difference between two eigenvalues then texture contains a unidirectional pattern and estimation of motion on both directions is not possible. If both eigenvalues are large and similar in scale, texture carries required information to extract motion in both directions.

T. Kanade and C.Tomasi used only the first criteria for feature selection. Their reasoning is given in [34] and is as follows. "In practice, when the smaller eigenvalue is sufficiently large to meet the noise criteria the matrix Z is also well conditioned. This is due to the fact that the intensity variations in a window are bounded by the maximum allowable pixel value, so that the greater eigenvalue can not be arbitrarily large".

If the eigenvalues for Z are (λ_1, λ_2) selection criteria for suitable regions can be thought as selecting regions with $\min(\lambda_1, \lambda_2) > \lambda_{thr}$ where λ_{thr} is a predefined threshold.

Window size is an important parameter effecting performance of the tracker. Selecting a large window would be advantageous for tracking large scale boundaries, similarly small scale edges can be tracked using small windows.

Eigenvalues threshold λ_{thr} is a parameter which determines the limit of horizontal and vertical edges for region selection. T. Kanade and C. Tomasi proposed to use average (λ_1, λ_2) value on uniform regions as a clue for selecting λ_{thr}. In this study window size selection is done using the same principal.

Several values for window size are compared for calculating (λ_1, λ_2) on face images. A proper window size should cause significantly higher (λ_1, λ_2) values on target face regions (eyebrows, eyes, mouth). This constraint brings an upper bound for window selection. If window size is greater than some value, averaging would cause eigenvalues on target regions to approach other facial regions which violates the first feature selection criteria.

The amount of displacement of target regions between two consecutive regions is another factor limiting window size. If window size is too small tracked region can completely disappear from window which will cause a misleading displacement calculation. A sample 480x640 face image (Figure 3.13) and eigenvalues for different window sizes are given in Figure 3.14.

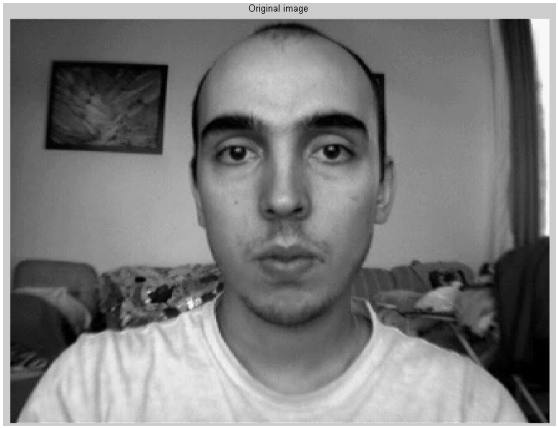


Figure 3-13 Sample Image Used for Window Size Selection

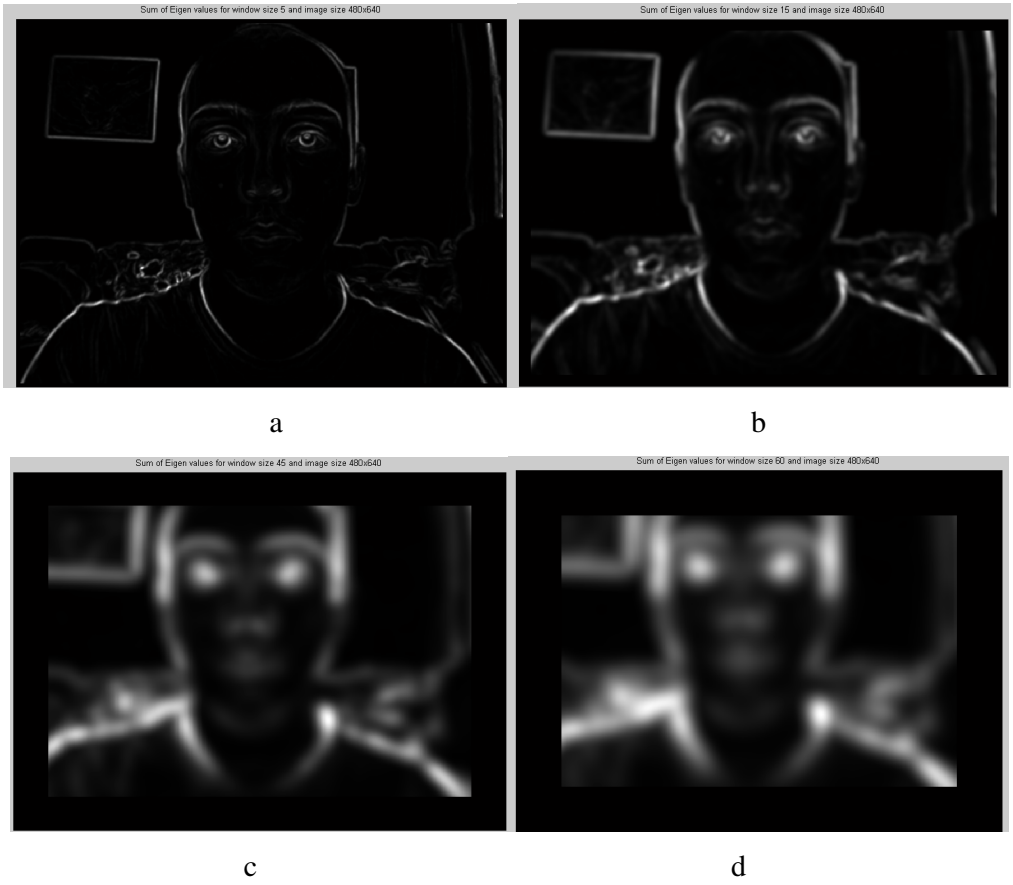


Figure 3-14 Eigenvalue Images for Different Window Sizes

Figure 3-14 shows the eigenvalues corresponding to window sizes 5x5(a), 15x15(b), 45x45(c) and 60x60(d). Target regions on first image have significantly larger eigenvalues but window size limits detectable displacements. Second image with a window size 15x15 still has relatively large eigenvalues on target regions and allows larger displacements to be detected. In the third and fourth images target regions start to mix with their surrounding which violates the first feature selection criteria. A window size selection of 15x15 allows the detection of large displacements between frames and target regions seem to carry significantly larger eigenvalues from their surrounding in Figure 3-14. Consequently window size 15x15 is chosen for KLT implementation.

3.4.3.Feature Tracking

In this study tracking of 18 feature points which are shown on Figure 3-11 are implemented using KLT tracker and FAFE feature extractor. Implemented technique is tested using sequences constructed from 640x480 frames.

Some of these feature points like eyebrows can be tracked reliably under several expressions however some other points like upper and lower boundary can shift from their true locations during tracking. The reason for this fact can be seen on Figure 3-14. In all 4 images eyebrows and eyes seem to have large eigenvalues which lead to successful tracking. However mouth regions have small eigenvalues which means a relatively uniform intensity and low displacement information on those regions.

Temporal occlusion or significant shape changes on feature regions also lead to feature losses during tracking. Some rules defining acceptable feature locations are defined and those rules are continuously checked during tracking. Lost feature points are detected using those defined rules. Flowchart for tracking algorithm implemented in this study is given on figure 3-15.

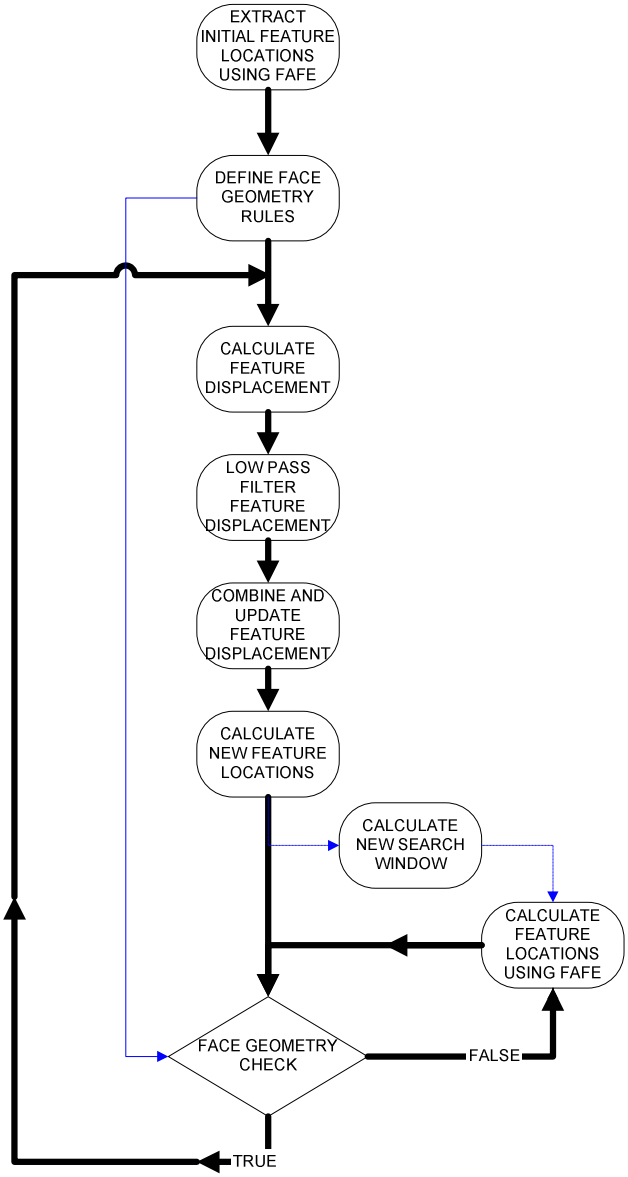


Figure 3-15 Feature Tracking Block Diagram

First step in feature tracking is detecting facial feature locations. Similar to the process described in section 3.3 feature localization is done using FAFE. Initial feature locations are assumed to be perfectly extracted and these coordinates are used for defining shape of the target face.

Feature displacements on the next frame are calculated using equation 3.34. Window size is selected as15x15 and a 2D Gaussian weighting function is used.

Feature displacements usually do not change instantly from frame to frame. Displacement calculation from the previous pair of frames is usually approximately equal to the displacement occurred between current frame pair. However dissimilar displacements which correspond to miscalculations caused by noise or occlusion may occur. Those unwanted peaky displacements can be removed by using a low pass filter. In this study a weighted averaging of current and previous frames is used to implement a low pass filter.

Let the solution of displacement equation 3.34 for 18 feature locations and frame index t is given by $D(t) = \{d_1(t), d_2(t), ..., d_{18}(t)\}$. The weighted averaging used for calculating effective displacements $D_{eff}(t)$ is given as follows.

$$D_{eff}(t) = 0.1(4D\ (t) + 3D\ (t-1) + 2D\ (t-2) + D\ (t-3)) \quad (3.35)$$

After low pass filtering current feature locations are updated using $D_{eff}(t)$.

Translations between expressions, natural movements like eye blinking or talking lead to shape changes on face. Although distances between several feature points can change significantly, others like distance between inner eye corners stay relatively similar in all cases as soon as face-camera distance does not change or target face does not rotate. In this study both of these constraints are assumed to be met.

As described before tracker can lose some or all of tracked features. This condition can be detected by continuously monitoring the validity of some distance restrictions on face. These restrictions are defined using feature pairs which stay in constant distances. Feature coordinates extracted from the first frame are used to calculate those constant distances and any significant variation from those values will be used as the sign of feature loss. Used feature location rules are described below.

Face can be divided to three regions in vertical direction corresponding to eyebrows, eyes, mouth. Those regions together with their centers and feature points on those regions are shown in Figure 3-15.

Let the distance between the features of index i and j is given by Δij and distance in the first frame is given by Θij. Distance between inner eye corners (feature points 9-11) is used as a threshold to detect unacceptable deviation from initial distances.

$$\tau = 0.25\Theta_{9-11} \tag{3.36}$$

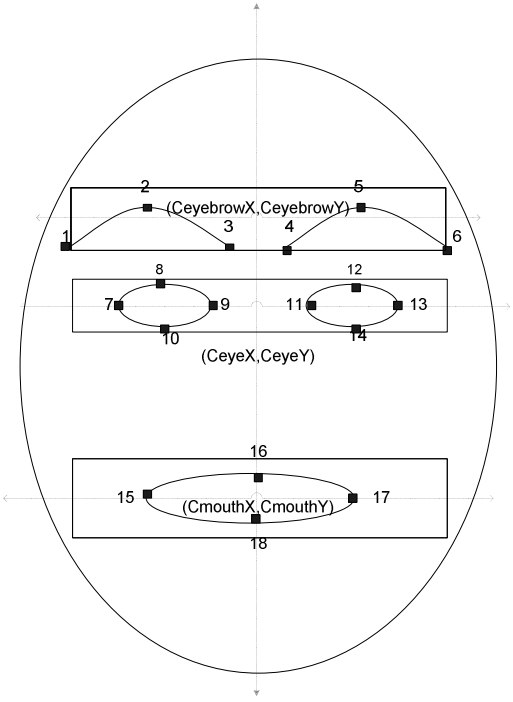


Figure 3-16 Face Template Used for Defining Face Geometry

1.) The first face geometry rule is horizontal alignment of three face regions centers which can be stated as follows:

$$\left|CeyebrowX - CeyeX\right| < \tau \;\&\; \left|CeyeX - CemouthX\right| < \tau \tag{3.37}$$

2.)The horizontal size of eyes is constant.

$$\left|\Delta_{7-9} - \Theta_{7-9}\right| < \tau \quad \& \quad \left|\Delta_{11-13} - \Theta_{11-13}\right| < \tau \tag{3.38}$$

3.)The vertical distance between center of eyes and top of mouth stay relatively constant.

$$\left|\Delta_{9-16} - \Theta_{9-16}\right| < 2\tau \quad \& \quad \left|\Delta_{11-16} - \Theta_{11-16}\right| < 2\tau \tag{3.39}$$

4.)If $(\vartheta xi, \vartheta yi)$ is the coordinate of feature point i then the following inequalities which define simple shape constraints hold for facial feature points

$$\begin{aligned} &\vartheta_{y7} - \tau < \vartheta_{y10} \quad \vartheta_{y9} - \tau < \vartheta_{y10} \quad \vartheta_{y8} - \tau < \vartheta_{y9} \quad \vartheta_{y8} - \tau < \vartheta_{y7} \\ &\vartheta_{y11} - \tau < \vartheta_{y14} \quad \vartheta_{y13} - \tau < \vartheta_{y14} \quad \vartheta_{y12} - \tau < \vartheta_{y13} \quad \vartheta_{y12} - \tau < \vartheta_{y11} \\ &\vartheta_{y15} - \tau < \vartheta_{y18} \quad \vartheta_{y17} - \tau < \vartheta_{y18} \quad \vartheta_{y12} - \tau < \vartheta_{y13} \quad \vartheta_{y12} - \tau < \vartheta_{y11} \end{aligned} \tag{3.40}$$

After each frame facial features are checked using the rules given. If there is a problem with those rules this is assumed as a feature loss and lost features are relocated using FAFE.

In order to decrease computation time of FAFE last valid feature locations are used as a clue to calculate COE and COM locations. Face localization and ellipse fitting operations are not repeated for recovering lost features. FAFE uses 10x10 search regions around last valid COE and COM coordinates to locate current COE and COM. Template parameters are recalculated once UFR, LFR templates are placed to COE and COM.

Loss of feature points could be caused by a long term occlusion of face and face can significantly drift from its last valid location during this period. If FAFE continuously tries to search the region defined by last valid COE and COM it may fail to detect correct locations although all features are clearly visible in current frame. In order to avoid this deadlock four consecutive invalid feature localizations are followed by a feature search on whole face.

3.5.Classification of Images Using Extracted Features

Feature Based Expression Classifier proposed in this study uses changes in feature distances between neutral and target image to classify it according to expression classes.

Let feature points extracted from a target image is given by $C_E = [p_0 \quad . \quad . \quad . \quad p_{18}]$ the distances between the feature pairs N feature pairs are calculated and they are stored into a matrix D_E. The same operation is repeated for corresponding neutral image and these values are stored in D_N .The distance changes between target image and neutral expression image is calculated:

$$\Delta D_E = (D_E - D_N) / D_E \tag{3.41}$$

The division by D_E term is included to count for possible scale differences between different images. Another effect of this term is to limit all entries in ΔD_E to the range [-1: 1].

ΔD_E is used as a feature vector to classify target image according to expression and named as facial actions (FA) in this study. The performances of different classifiers (Bayes, NM, KNN, NN) are compared using AF Face database.

The idea used in Bayes classification can be written as follows. Assume face view with a neutral expression. Later we observed a set of facial actions $\Delta D_E = \{\Delta D_{E1},...,\Delta D_{EN}\}$ and face contains an expression after these changes.

Probability of having an expression E_i is given as $P(\mathrm{E}_i \mid \Delta D_E)$. Using the basic relation of probability we can write $P(\mathrm{E}_i \mid \Delta D_E).P(\Delta D_E) = P(\Delta D_E \mid \mathrm{E}_i).P(\mathrm{E}_i)$

and if we rearrange this equation:

$$P(\mathrm{E}_i \mid \Delta D_E). = \frac{P(\Delta D_E \mid \mathrm{E}_i).P(\mathrm{E}_i)}{P(\Delta D_E)} \tag{3.42}$$

If we want to make a decision on resultant expression over a set of possible expressions $\{E_1,....,E_M\}$ we can use Bayes decision rule which is based on choosing the expression with maximum $P(E_i \mid \Delta D_E)$. Since $P(\Delta D_E)$ is common for each expression we can remove this term. Similarly if we assume a identical priori probability for each expression Bayes decision rule turns to:
Decide on expression E_i with $P(\Delta D_E \mid E_i) > P(\Delta D_E \mid E_{j \neq i})$

In this study the class conditional probability of each facial action is assumed as independent and using this assumption $P(\Delta D_E \mid E_i)$ is calculated using the following equality.

$$P(\Delta D_E \mid E_i) = \prod_{x=\{1,...,N\}} P(\Delta D_{Ex} \mid E_i) \tag{3.43}$$

Each probability density $P(\Delta D_{Ex} \mid E_i)$ is approximated using histogram of facial actions taken in training expression – neutral image pairs. Calculated histograms are fit to Gaussian distribution and used for classification. Results obtained are described in chapter 5.

CHAPTER 4

GABOR WAVELET BASED APPROACH

Feature Based Technique proposed in previous chapters is based on extraction of facial features from images and analysis of changes in these feature locations. In this chapter an alternative holistic method which analyzes face as a whole without partitioning is described.

Gabor filters which found several applications in literature are used for extracting local edge information at various scales and orientations. Filtered images (Gabor jets) are later classified according to expression on the image. Dimension of extracted Gabor jets are decreased using PCA and the effect of this conversion on classifier performance is analyzed.

First some information about Gabor Wavelet Function will be presented in section 4.1. Some examples to the usage of Gabor filters in human face analyses will be given in section 4.2. The usage of Gabor filters in this study is described in section 4.3. Distance measures commonly used for Gabor jets are described in section 4.4 and usage of Gabor jets for expression classification is described in section 4.5.

4.1.Gabor Wavelet Functions

There has been an increasing trend in image processing to represent image data using wavelets which provides multi-scale and multi-orientation representations. This trend results from the fact that features of interest are usually available at various scales and angles. Gabor wavelet functions are widely used to analyze such features.

Gabor wavelet functions are Gaussians modulated by complex sinusoids. The general form of 2D Gabor wavelets and their Fourier transform is given in equations 4.1 and 4.2 respectively. Gabor function plots for (σ_x, σ_y)=(10,20) and (u_0, v_0)=(0.2,-0.1) are given in figures 4.3 to 4.5.

$$\varphi(x, y; u_0, v_0, \sigma_x, \sigma_y) = e^{(-\pi(\frac{x^2}{2\sigma_x^2}+\frac{y^2}{2\sigma_y^2})+2\pi i(u_0 x+v_0 y))} \tag{4.1}$$

$$\Psi(u, v; u_0, v_0, \sigma_x, \sigma_y) = e^{(-2\pi^2(\sigma_x^2(u-u_0)^2+\sigma_y^2(v-v_0)^2)} \tag{4.2}$$

σ_x , σ_y :Width of Gaussian in spatial domain

u_0 , v_0 :Frequency of complex sinusoid

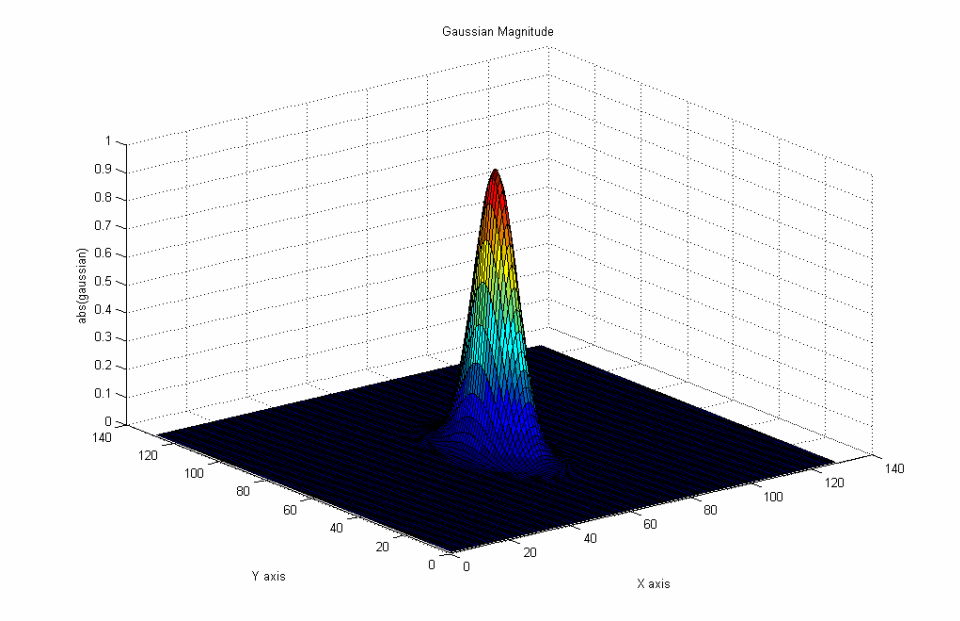


Figure 4-1 Modulated Gaussian

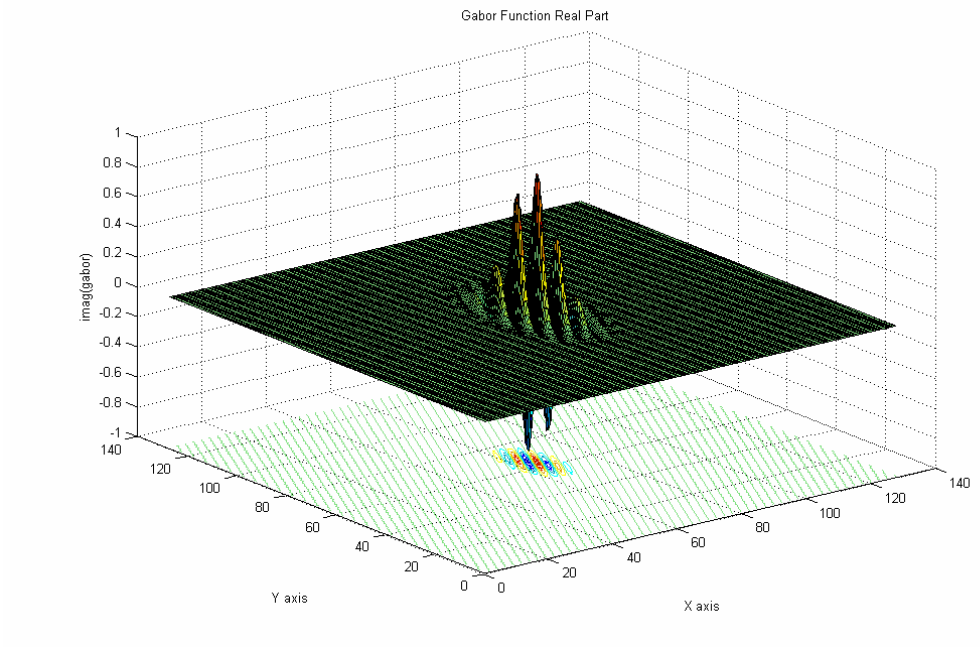


Figure 4-2 Real part of Gabor Function

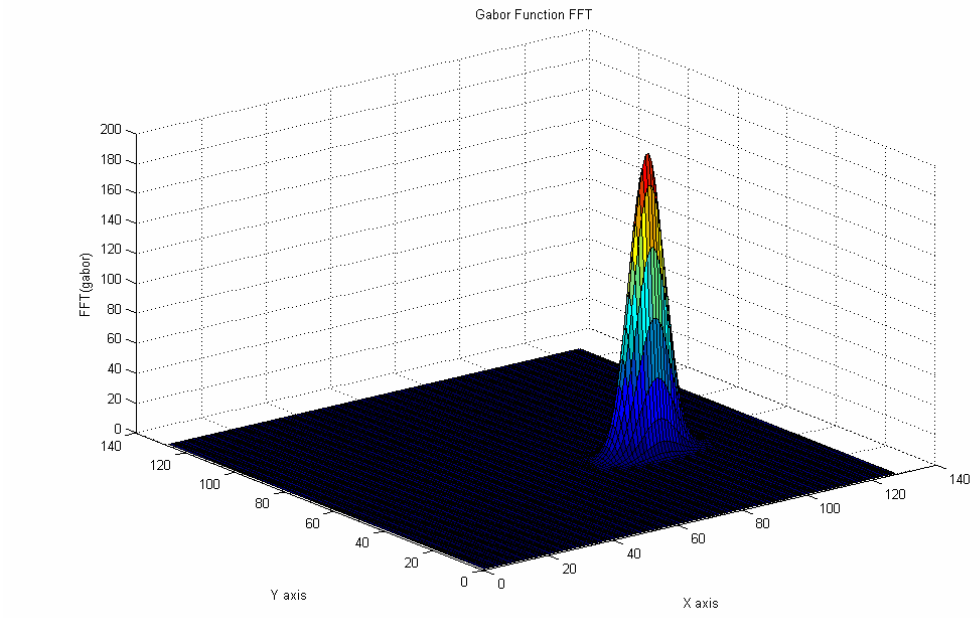


Figure 4-3 2D FFT of Gabor Function

Uncertainty principle which can be stated as $\Delta f \Delta t \geq 1/4\pi$ where Δf is the frequency resolution and Δt is the time resolution brings a lower bound on the product of resolution in time and frequency which is defined as joint time frequency resolution (JTFR). J.G.Daugman [13] has shown that Gabor functions are unique in attaining the minimum possible value for JTFR.

Another important property of Gabor functions is their similarity to mammal visual system with their frequency and orientation representation characteristics. Marcelja (1980) and Daugman (1983) suggested that the receptive fields (RF) of cortical simple cells might be well described by a Gabor function. Studies of Jones and Palmer showed that a general 2D Gabor function can very closely match the RF profile of simple cells.

Jones and Palmer tested Daugman's and Marcelja's hypothesis which divides simple receptive fields into a class of linear spatial filters analogous to 2D Gabor filters. In the space domain, they found 2D Gabor filters that fit the 2D spatial response profile of each simple cell in the least-squared error sense. They also found a Gabor function that fit the 2D spectral response profile of each simple cell and observed that the residual errors were minimal and unstructured. They conclude that the Gabor function provides a useful and reasonably accurate description of most spatial aspects of simple receptive fields [14].

Although there have been very evolutionary advances in machine vision algorithms in the last few decades, several tasks very easily done by human vision system are still quite challenging for machines. One straight forward methodology to solve this lack of performance is to enumerate biology in machine vision applications.

Gabor filter which has similar characteristics to mammal vision system has performed quite well in very diverse fields like edge detection, texture analysis, motion estimation and image compression. Human face analysis is also another important field where Gabor filters are widely used and more about this will be described in the following sections.

4.2.Gabor Filters In Human Face Analysis

Increasing importance of HMI systems in our daily life has led to more studies on human face. Face recognition, expression recognition, eye tracking, lip reading, driver surveillance systems are some examples that can be given to studies on human face, all of which can be named as facial analysis. Although the definition of problem is quite different in these studies, required features for the solution and image processing techniques used are very similar.

Face can be thought as a complex structure that is composed of several independent regions like facial organs and facial hairs. All these independent regions can show great change in their shape. Some of these changes like opening eyes and rising eyebrows are correlated with each other while others like opening mouth and closing eyes are uncorrelated.

Although most important patterns on face appear on facial organs and facial hairs, temporal or constant wrinkles on skin are also important clues that can be used for facial analysis. As an example facial patterns around eyes and mouth represent very important features for expression classifications.

Analysis of facial patterns can be thought as a problem of detecting features from various scales and orientations. Local illumination differences around these patterns are also an important issue which can cause false parameterization of facial patterns.

Gabor wavelet filters with their wide usage in image processing applications are shown to be reliable detectors for features with varying angle and size. They are also robust to brightness variations. These two properties make Gabor filter based feature detection a reasonable choice for facial pattern analysis. In most facial analysis studies the aim is to emulate human understanding of facial scenes. Human understanding about a facial scene can be divided into two steps which are capturing the information from sensory organs and processing this information to make a decision on the scene. We have very little information about the processing of data in human brain but we know the similarity between sensory organ

characteristics and Gabor wavelet filtering characteristics which makes Gabor filtering a very important feature extraction method for facial analysis.

Gabor filter responses are very powerful features for face analysis problems and are widely used for the last two decades. Studies of M.Lades [16] (1993) are among the first studies using Gabor wavelet functions for the analysis of facial images. They used Gabor filter responses and elastic graph matching to recognize facial images in gray level images. Their study was also the first study using equation 4.2 as a model for Gabor wavelet functions.

Z. Zhang [17] (1998) compared two types of features to detect facial expressions one of which was the geometric positions of 34 judicial points and the other was 612 Gabor wavelet coefficients. M.J.Lyons [18] (1999) combined Gabor wavelet analysis and elastic graph matching with discriminant analysis to classify facial images according to race age and expression classes. X.Cao [19] used Gabor wavelet filters to extract parameters of a face avatar and proposed the usage of this avatar for low bandwidth implementation of video conferencing. I Bucio [20] compared the performance of different classifiers and different distance measures to classify images according to expressions using Gabor filtered images. B Gökberk [21] analyzed discriminative power of feature locations on Gabor filtered images for optimal face recognition and showed most discriminative information is carried by region around eyes while nose and hair region have large variations. Q.Ji [23] combined infrared based eye tracking with Gabor feature detection and tracking and showed facial features can reliably tracked under very rapid head motion and instant expression changes.

The most widely used form of Gabor filters in facial analysis is a modified version of general Gabor function given in equation 4.1. This modified version uses modulated Gaussian with identical width in both vertical and horizontal axis and the modulating planar sinusoidal frequencies are represented in polar form. The modified version of equation 4.1 is given in equation 4.3. If z is the coordinate vector z=(x, y)

$$\psi_k(z) = \frac{kk^T}{\sigma^2}\exp(-\frac{kk^T}{2\sigma^2}z^T z)(\exp(ik^T z) - \exp(-\frac{\sigma^2}{2})) \tag{4.3}$$

$$k = (k_v \cos\varphi_\mu \qquad k_v \sin\varphi_\mu)^T \tag{4.4}$$

$$k_v = 2^{-\frac{v+2}{2}}\pi \tag{4.5}$$

$$\varphi_\mu = \mu\frac{\pi}{8} \tag{4.6}$$

The Gabor wavelet representation in equation 4.3 allows description of spatial frequency structure in image while preserving information about spatial relations The second exponential term on equation 4.3 defines the oscillatory part. The term $\exp(-\frac{\sigma^2}{2})$ is subtracted to render the filters insensitive to the overall illumination. Equation 4.3 does not normalize energy picked up by a kernel in convolution. D Field [15] showed that the power spectrum of natural images decrease by $\frac{1}{\|k\|^2}$ so $\frac{kk^T}{\sigma^2}$ is used to ensure that filters tuned to different spatial frequencies have approximately equal energy.

The amplitudes of the filter responses are used as features to test for the presence of spatial structure restricted to a band of orientations and spatial frequencies. The Gaussian envelope given by $\exp(-\frac{kk^T}{2\sigma^2}z^T z)$ assures that the amplitude information degrades gracefully with shifts from the location where it is sampled and allows spatial analysis.

Sharp edges are important features for object recognition. Gabor filters are known to respond strongly to edges if they are perpendicular to the direction of $\vec{k}$. However real and imaginary parts of the filter response oscillate with the characteristic frequency of the filter in place of providing smooth peaks for edge detection.

One solution to this problem is to abandon the linearity of the transform and use the magnitude of the filter response [16] and in this study only the magnitude of Gabor filter responses have been used. Equation 4.7 gives the Gabor wavelet filter response of an image I .

$$WI(\vec{k},\vec{x}_0) = \iint \psi_k(\vec{x}-\vec{x}_0) I(\vec{x}) d^2\vec{x} = (\psi_k * I)(\vec{x}_0) \tag{4.7}$$

$\vec{x}_0$: Center location of Gabor wavelet filter

σ specifies the radius of the Gaussian. Following wavelet terminology the size of the Gaussian can also be called as the wavelet's basis of support. The Gaussian size determines the amount of the image that effects convolution. As the convolution moves further from the center of the Gaussian, the remaining computation becomes negligible and the significant part of the convolution comes from the region close to the Gaussian center. Figure 4.4 shows Gabor filter impulse responses for $\sigma = \{\frac{\pi}{2}, \pi, \frac{3\pi}{2}, 2\pi\}$ and $v = 4$, $\mu = 4$

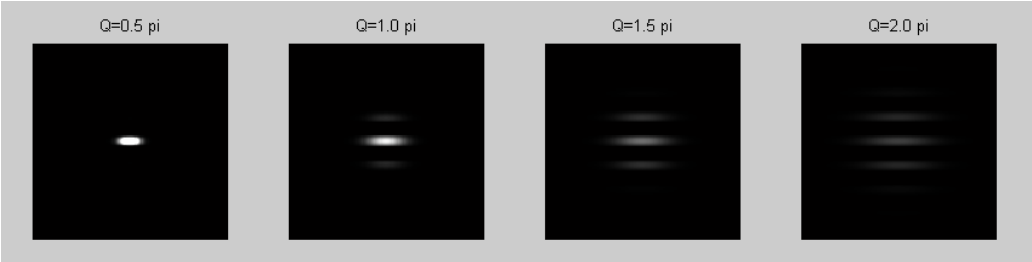


Figure 4-4 Effect of Gaussian Radius on Gabor Filter Impulse Responses

v specifies the frequency of the plane wave, Gabor wavelets with a high frequency will respond to sharp edges and bars while wavelets with a low frequency will respond gradual changes in intensity in the image. Figure 4.5 shows Gabor filter impulse responses for $v = \{1,2,4,8\}$ and $\sigma = \pi$, $\mu = 4$

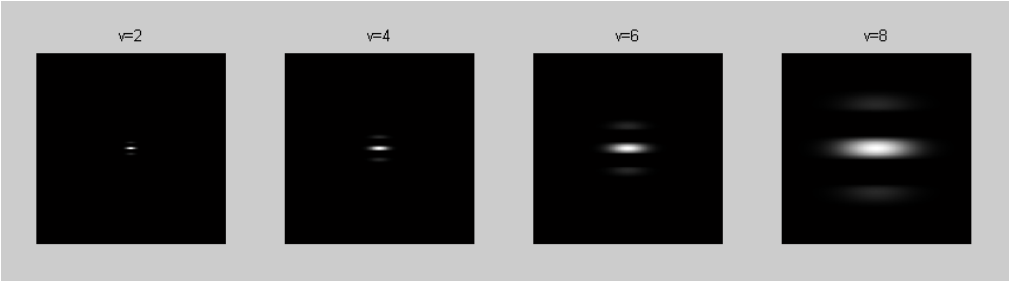


Figure 4-5 Effect of Modulating Frequency on Gabor Filter Impulse Responses

μ specifies the orientation of the wavelet. This parameter rotates the wavelet about its center. The orientation of the wavelets dictates the angle of the edges or bars for which the wavelet will respond. Wavelets will give high responses on regions with edges perpendicular or nearly perpendicular to their angle.

In most cases μ is a set of values from 1 to 8 corresponding to angles 0 to π. Values from π to 2π are redundant due to the symmetry of the wavelet response magnitude. For example consider two wavelets with an angular distance of π, the imaginary and real parts of wavelets will change sign. This change results in a response with opposite sign, but since we are only interested with the magnitude of the response those two filters would be identical. Figure 4.6 shows Gabor wavelets from 8 orientations with $\mu = \{1,2,...,8\}$ and $\sigma = \pi, v = 4$

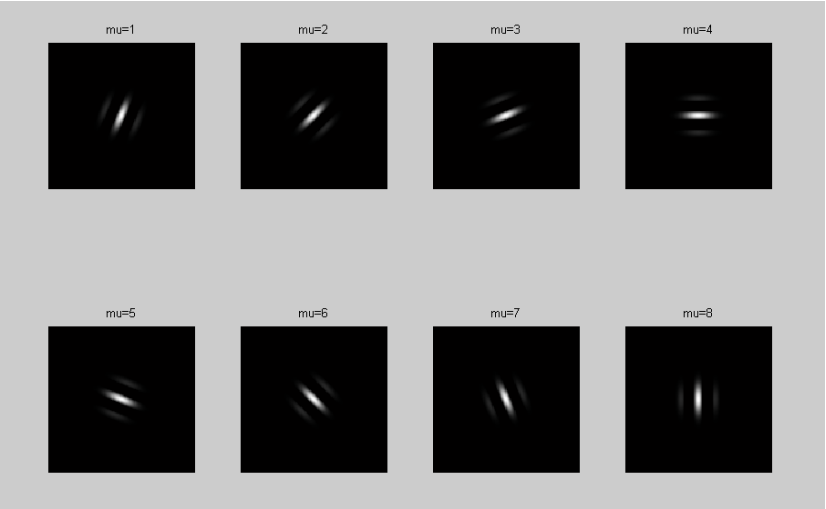


Figure 4-6 Effect of orientation on Gabor Filter Impulse Responses

The frequency responses of filters are directly related to the angle, width and frequency parameters of the Gabor function. In fact they show a bandpass characteristics with angle defined by the μ, frequency selectivity defined by Gaussian width σ and distance of the filter center defined by v. This fact is shown in figures 4.7 to 4.9.

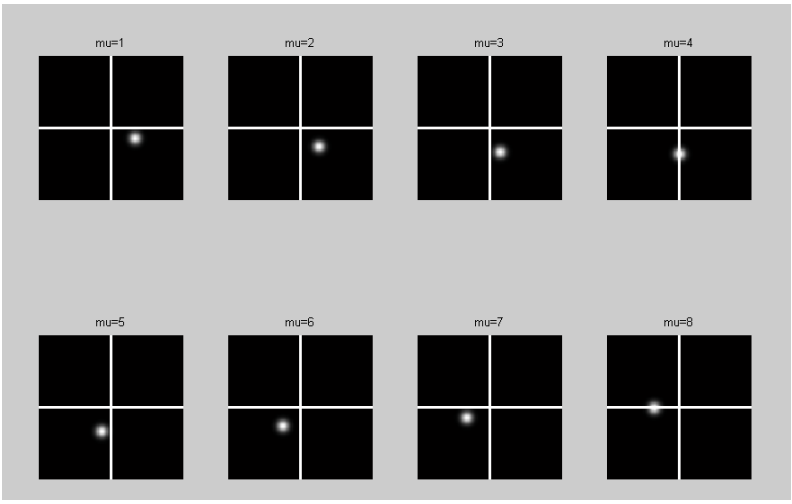


Figure 4-7 FFT of Gabor filters for $\mu = \{1,2,...,8\}$ and $\sigma = 2\pi$, $v = 4$

Change of frequency response with the change of orientation control parameter μ is given in figure 4.7. As can clearly be seen from the figure changing values of μ directly effects the location of bandpass filter and frequencies corresponding to the angle defined by μ are selected by the filter.

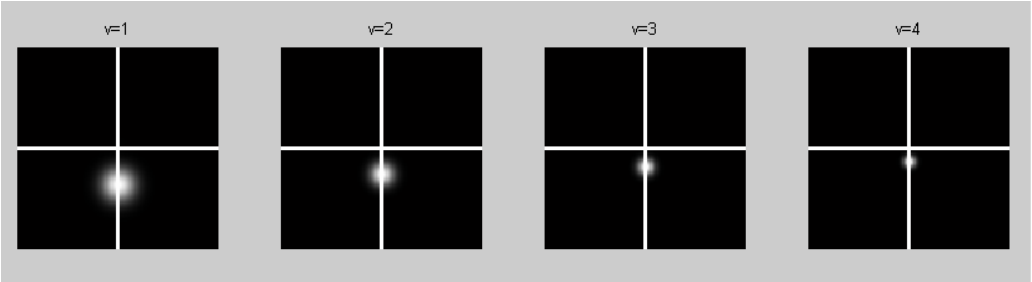


Figure 4-8 FFT of Gabor filters for $\mu = 4$ and $\sigma = \pi$, $v = \{1,2,3,4\}$

Changing values of v increases the frequency of modulating plane wave, shifting the frequency response of Gabor filter. The first exponential term in equation 4.2 defines the modulated Gaussian whose width is defined by $\frac{k}{\sigma}$. Decreasing k (increasing v) results in an increase in filter width, this effect can clearly be seen in figure 4.8. Because of the combined effect of v on filter center frequency and filter width usally a constant σ is selected The change of modulated Gaussian width σ effects the bandwidth of Gabor filter. Filters corresponding to large values of σ decay slowly in time domain and show a small bandwidth characteristic in frequency domain, this can be seen in figure 4.9.

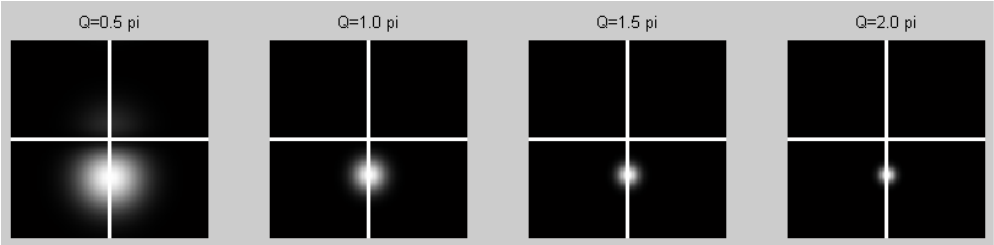


Figure 4-9 FFT of Gabor filters for $\mu = 4$ and $\sigma = \{\frac{1}{2}\pi, \pi, \frac{3}{2}\pi, 2\pi\}, v = 1$

4.3.Gabor Jet Face Representation

Gabor wavelet filters are tuned for the detection of edges in a range of frequencies and a range of orientations. Face is a structure containing several temporal or permanent boundaries at several orientations. The scale of those boundaries differs significantly both between each other and between different people. Any analysis on face should count for all those boundaries which give highly valuable clues about face like identity or expression. One solution to this problem is to use multiple filters tuned to different orientations and scales. If $I(z)$ represent the target image and $\psi_{v,\mu,\sigma}(z)$ is a Gabor filter tuned with parameters (v, μ, σ) then the Gabor filtered image $J_{v,\mu,\sigma}(z)$ can be written as.

$$J_{v,\mu,\sigma}(z) = \int I(z^{'}) \psi_{v,\mu,\sigma}(z - z^{'}) d^2 z \tag{4.8}$$

The feature used for Gabor filter image analysis in this study is a concatenation of filtered images $J_{v,\mu,\sigma}(z)$ with different tuning parameters. This concatenated set is called a Jet which can be mathematically represented as

$$Jet(z) = \begin{bmatrix} J_{v1,\mu1,\sigma1}(z) \\ J_{v2,\mu2,\sigma2}(z) \\ . \\ J_{vn,\mu n,\sigma n}(z) \end{bmatrix}_{\{v1,\ldots,vn\} \in V \quad ,\{\mu1,\ldots,\mu n\} \in \mu \quad ,\{\sigma1,\ldots,\sigma n\} \in \Omega} \tag{4.9}$$

The range choice for frequency, Gaussian radius and orientation parameters differ in the literature. Different choices of these parameters for some face analysis studies in the literature are given on table 4.1.

Table 4.1 Gabor Filter Parameter Ranges for Some Studies

Reference #	v Range	μ Range	σ Range	Image Size
X.Cao [19]	0,…,3	0,…,5	π	256x256
I.Bucio [20]	0,…,4	0,…,3	π	80x60
M.J.Lyons[18]	1,…,5	0,…,5	π	NA
L Wiskott [22]	0,…,4	0,…,7	2π	NA
H.Gu Zhu[23]	1,2,4	0,…,5	π	128x128

(NA: Information not available)

In this study Gabor jets extracted from 60x45 images are used as features to make classification according to expressions. Different ranges for v, μ are evaluated and their performances are compared. Success rates for different ranges will be given on chapter 5. The highest success rate was reached using the ranges given below.

$\sigma = \pi$ v= {0,…,5} μ ={0,…,5}

Total size of face jet for a given 60x45 image reaches 97200. Figure 4.10 shows a sample image used to show the effect of Gabor filtering. Figure 4.11 shows the real parts of used Gabor filter impulse responses. Figure 4.12 shows the magnitudes of filtered images.

Figure 4-10 Sample Image Used for Gabor Jet Calculation

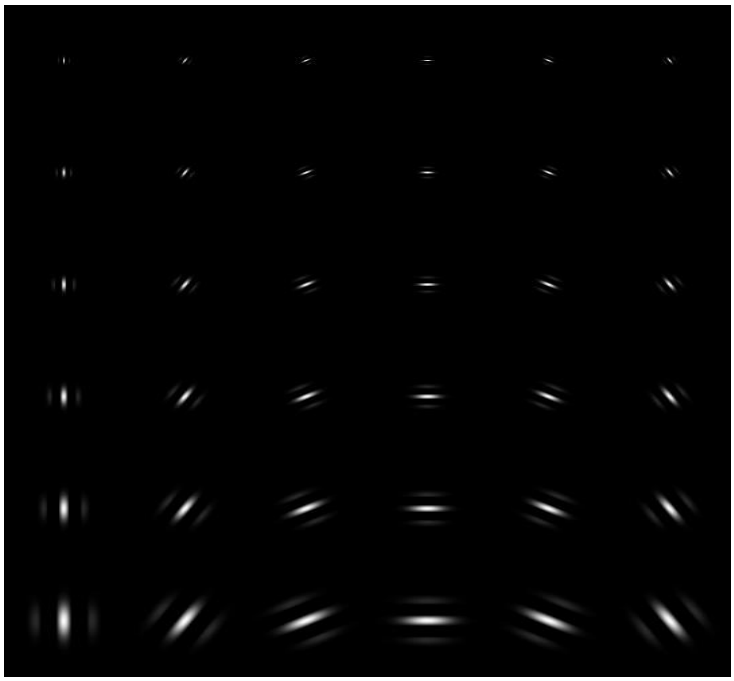


Figure 4-11 Used Gabor Filter Impulse Responses

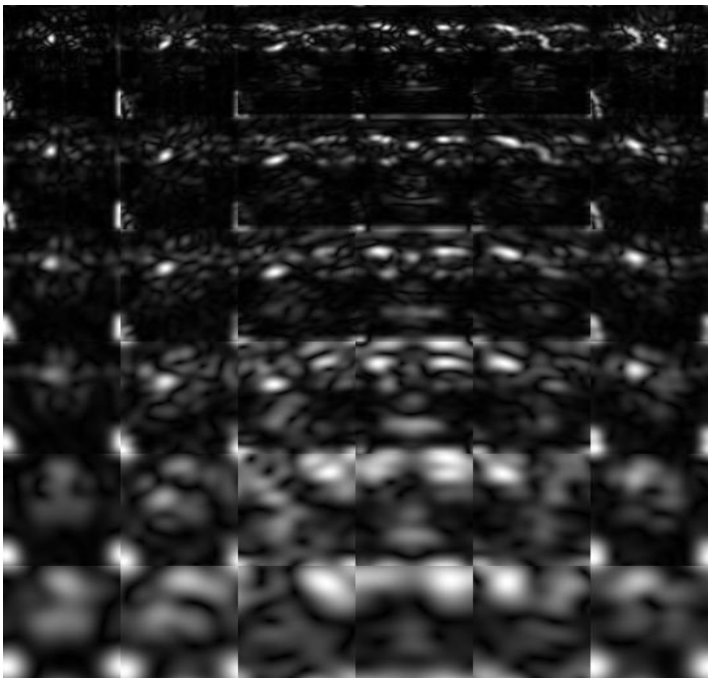


Figure 4-12 Gabor Filtered Forms of Sample Image

Gabor filter set used in this study only analyzes a range of frequencies and orientations. The sum of used Gabor filter responses in frequency domain is given in figure 4.15. As can be seen from the figure, used filter set occupies half of the possible orientations (which means all orientations are analyzed since we only use the magnitude of filter response) and about one third of frequency range.

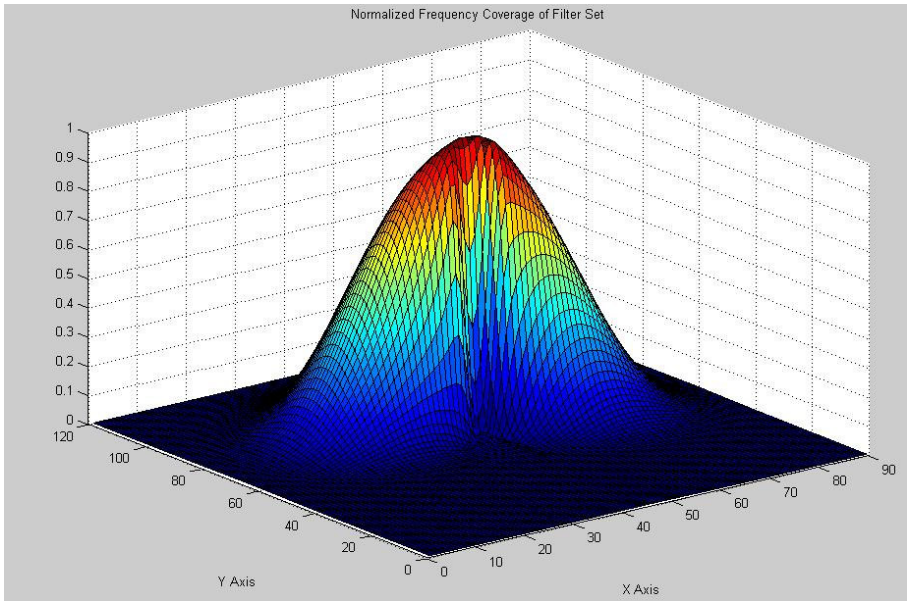


Figure 4-13 Normalized Frequency Coverage of Gabor Filter Set

4.4.Distance Measures for Gabor Jets

The complex valued Gabor jets can be written in polar form as given in equation 4.10 with magnitude given by $A(z)$ and phase given by $\theta(z)$. The magnitude information smoothly changes while the phase information varies approximately with the characteristic frequency of the respective Gabor filter [22]. Small spatial displacements cause significant variation in phase and due to this variation two jets cannot be directly compared.

$$J_{v,\mu,\sigma}(z) = A(z)\exp(i\theta(z)) \tag{4.10}$$

A straightforward solution is to use only the magnitude information. When only the magnitude information used image jet gets the form of a real 2D matrix. This matrix can also be turned to a one dimensional matrix simply by changing its shape without loosing any information.

Most popular distance measures used for feature sets in pattern recognition are Euclidian (EDM), Cosine (CDM), Mahalanobis (MDM) distance measures. Cosine distance measure is widely used for comparison of two Gabor jets and uses the angle between two jets as a distance measure. CDM captures a scale invariant understanding of similarity since independent of vector size and magnitude of its entries it does always give a result between -1 and 1. CDM for two jets $Jet_1(z)$ and $Jet_2(z)$ is given in equation 4.11

$$CDM_{1-2} = \frac{Jet_1(z)^T Jet_2(z)}{\|Jet_1(z)\|\|Jet_2(z)\|} \tag{4.11}$$

Euclidian distance can be defined as the length of line connecting two points in an N Dimensional space. Euclidian distance between two jets is given in equation 4.12.

$$\mathrm{EDM}_{1-2} = \|Jet_1(z) - Jet_2(z)\| \tag{4.12}$$

Most significant weakness of Euclidean distance is its dependence to the scale of vectors compared. This distance measure also does not take into account the correlation between vector entries. Solution to this weaknesses is a distance measure named as Mahalanobis distance introduced by P. C. Mahalanobis in 1936.

It differs from Euclidean distance in that it takes into account the correlations of the data set. Mahalanobis distance is given in equation 4.13.

$$\mathrm{MDM}_{1-2} = \sqrt{(Jet_1(z) - Jet_2(z))^T \Sigma^{-1} (Jet_1(z) - Jet_2(z))} \tag{4.13}$$

Σ : Covariance matrix of a training Jet set.

The diagonal entries in Σ are directly related with the scale of jet matrix entries and the inverse of Σ makes an effect of normalizing jet entries and making this distance measure scale invariant.

4.5. Expression Classification Using Image Jets

Different from feature based classifier, expression classifier based on Gabor jets does not require knowledge about the neutral expression image. In this technique similar to other expressions neutral expression is used as an expression class used for classification.

Face is not partitioned to regions and Gabor jets calculated from the whole face are used for classification. Performances of different classifiers (NN, KNN, NM) and different distance measures are compared using AF and JAFFE databases.

Used classifiers are based on calculating distance between target image jet and training jets. The computation time especially for KNN and NM classifiers grow rapidly with training database size and jet size. In order to solve this problem image jets are projected to a lower dimensional subspace using PCA and the effect of this conversion to classifier performances is evaluated. Results are given in chapter 5.

CHAPTER 5

RESULTS AND PERFORMANCE ANALYSIS

Two techniques (Feature Based, Holistic) for facial expression on still images are implemented in this study. Their performances are compared using two facial expression databases (JAFFE, AF). Information about these databases is presented in section 5.1. A fully automatic facial feature extractor (FAFE) is proposed and its performance is evaluated using hand marked feature coordinates. Obtained results are described in section 5.2. A facial feature tracking technique using KLT tracker and FAFE is implemented and its performance is evaluated using sample sequences, comments on obtained results are presented in section 5.3. Proposed feature extractor is used for coding face images from AF database into facial distance vectors which show distances between predefined facial feature points. Changes of facial distances between test image (containing an expression) and corresponding neutral expression image are used for expression classification. Classification performances of different classifiers are presented in section 5.4. Holistic facial expression classifier is tested using JAFFE and AF databases. Results obtained using different classifiers and different distance measures are presented in section 5.5

5.1.Used Databases

Two databases are used for measuring the performances of implemented classifiers. JAFFE (Japanese Female Facial Expression) database is constructed using 213 face images of 7 expressions (Neutral, Surprise, Sad, Happy, Fear, Disgust, Anger) taken from 10 Japanese females. Images are given in gray scale and contain an over shoulder view of a person. Image size for this database is 256x256. For each person

images containing all 7 expressions are captured in multiple sessions. Some sample pictures from this database are given in Figure 5.1.

Figure 5-1 Sample Pictures from JAFFE Database

AF (Aleix Face) database is constructed from 576x768 color images of 126 people (70 men and 56 women). No restrictions on wear (clothes, glasses, etc.), make-up, hair style, etc. were imposed on participants. Images are taken under 13 conditions including 4 expressions, different illumination and different occlusion conditions (sun glasses, scarf). Images are captured in 2 sessions two weeks apart. In this study only images corresponding to 4 expressions (Neutral, Smile, Anger, and Scream) are used. Some sample pictures from the database are given in Figure 5.2.

Figure 5-2 Sample Pictures from AF Database

AF database is divided into training (344 images) and test (320 images) sets. 18 feature coordinates corresponding to UFR and LFR regions are hand marked. Hand marked feature coordinates in training database are used both for principal component calculations and classifier training. Hand marked information on test set is used for performance analysis of FAFE feature extractor.

The limited subject and image count in JAFFE database does not permit division of this database into training and test sets. Holistic expression classifier performance on this database is calculated using leave one out strategy. This means usage of all images except the analyzed one as training patterns and repeating training phase for each tested image.

5.2.FAFE Feature Extractor Performance

Performance of FAFE feature extractor is measured using AF database. On each image the distances between hand marked and extracted feature coordinates are measured.

The most important parameter effecting feature localization performance is the number of principal components (PCC) which are used both for template localization and template parameter calculations (In this study the number of principal components used for both template localization and template parameter calculations are identical). Change in the feature extraction performance with varying PCC is given in Table 5-1. In calculation of feature extraction error deviations greater than 50 pixels are assumed to be caused by incorrect template localization and they are ignored.

Increasing the number of principal components increases the span of their subspace. A larger subspace contains closer members to upper and lower facial regions. This should increase both template localization and template matching performance of FAFE. The results obtained (shown in Table-5-1) are consistent with this idea.

As the number of principal components increase, performance of feature extractor clearly increases until the number of principal components reaches around 16 and stays in a narrow range for higher number of components. Although FAFE performance for 30 principal components is better, performance obtained using 16 components is found to be acceptable and computation time is limited using a subspace with a dimensionality of 16.

Table 5-1 Feature Extraction Error (Pixels) with Changing PCC Number of Principal Components (Bold Entries Show Minimum Error for Each Feature)

Feature #	1	2	4	8	16	30
1	11,00	11,01	10,72	10,25	**9,96**	10,17
2	10,08	9,95	9,74	9,50	**9,34**	9,47
3	8,89	8,83	8,57	8,48	8,43	**8,34**
4	8,45	8,65	8,40	8,32	8,07	**7,82**
5	9,38	9,38	9,03	8,76	8,72	**8,50**
6	10,57	10,82	10,57	10,13	10,09	**10,01**
7	8,77	8,62	8,38	8,07	8,05	**7,83**
8	8,73	8,82	8,52	8,12	7,98	**7,75**
9	7,83	7,91	7,60	7,54	7,35	**7,19**
10	8,26	8,35	7,86	7,87	7,59	**7,34**
11	7,94	7,80	7,40	7,35	7,30	**7,25**
12	8,71	8,61	8,29	8,11	7,87	**7,71**
13	9,63	9,57	9,40	9,20	9,00	**8,64**
14	8,64	8,57	8,32	8,20	8,01	**7,65**
15	9,78	8,31	7,78	7,71	**6,65**	7,09
16	10,64	9,89	7,77	7,85	**6,98**	7,07
17	10,86	9,23	8,51	8,34	**6,72**	7,05
18	14,71	12,42	9,51	**9,10**	9,94	9,96

5.3.Feature Tracking Performance Using KLT and FAFE

Another usage of FAFE in this work is initialization and recovery of lost facial features in KLT tracker. Some rules are defined to describe the face geometry and those rules are continuously checked during tracking. If any violations are detected feature coordinates are recovered using FAFE. Performance of this system is tested using sample video sequences.

Feature point corresponding to lower mouth boundary is observed to be the hardest point to detect and track. This is due to the weak horizontal boundary on this region which has approximately similar intensity regions on both sides.

Images with widely open mouths also result in false localization of lower mouth. This problem can be described with the large shape variation between open, closed and widely open mouth shapes which does not permit the modeling of this region with used principal subspace. Upper mouth and lower mouth feature locations (with index 16, 18) which have low vertical edge information have a tendency to shift in horizontal directions.

Regions corresponding to upper and lower eyelid boundaries combine during eye blinking. In some cases this leads to replacement of regions during tracking. Relative positions of these two regions are monitored and relocated in case they are lost. Occlusion of face by hand or rotation leads to the loss of tracked features. Video sequences containing frames with partially occluded face are used for testing the performance of tracker under such conditions.

An example sequence is given in Figure 5-3. With appearance of hand (frame B), mouth region gets occluded and feature corresponding to lower mouth boundary moves from its correct location. Complete occlusion of mouth (frame C) cause feature locations to violate the eye-mouth alignment rule and features are relocated using FAFE (frame E).

Figure 5-3 Sample Sequence of Partial Occlusion (Frames with a detected face geometry rule violation contain a red rectangle in upper left corner)

5.4.Feature Based Expression Classifier Performance

Facial actions calculated using feature extracted images are used as features for different classifiers. Details about facial action (FA) calculations are presented in section 3.5. A set of facial feature pairs is selected for defining facial distances. Changes in these distances between target image and corresponding neutral expression image are used for calculating facial actions which correspond to the rate of decrease or increase in facial distances. A set of facial actions with corresponding facial features (Figure 3-16 shows location of each facial feature) is given in Table 5.2.

Table 5-2 Facial Actions Used by Feature Based Classifier

	FA1	**FA2**	**FA3**	**FA4**	**FA5**	**FA6**
FEATURE INDEXES	2-8	8-10	16-15	15-17	15-10	15-18

Bayes classifier requires class conditional densities for each FA. Density functions for three expressions are calculated using histogram of each FA. Histograms of six FA for used expressions are given in Figure5-4. Calculated histogram means and variations are used for finding best fitting Gaussians which are used as class conditional density functions for facial actions.

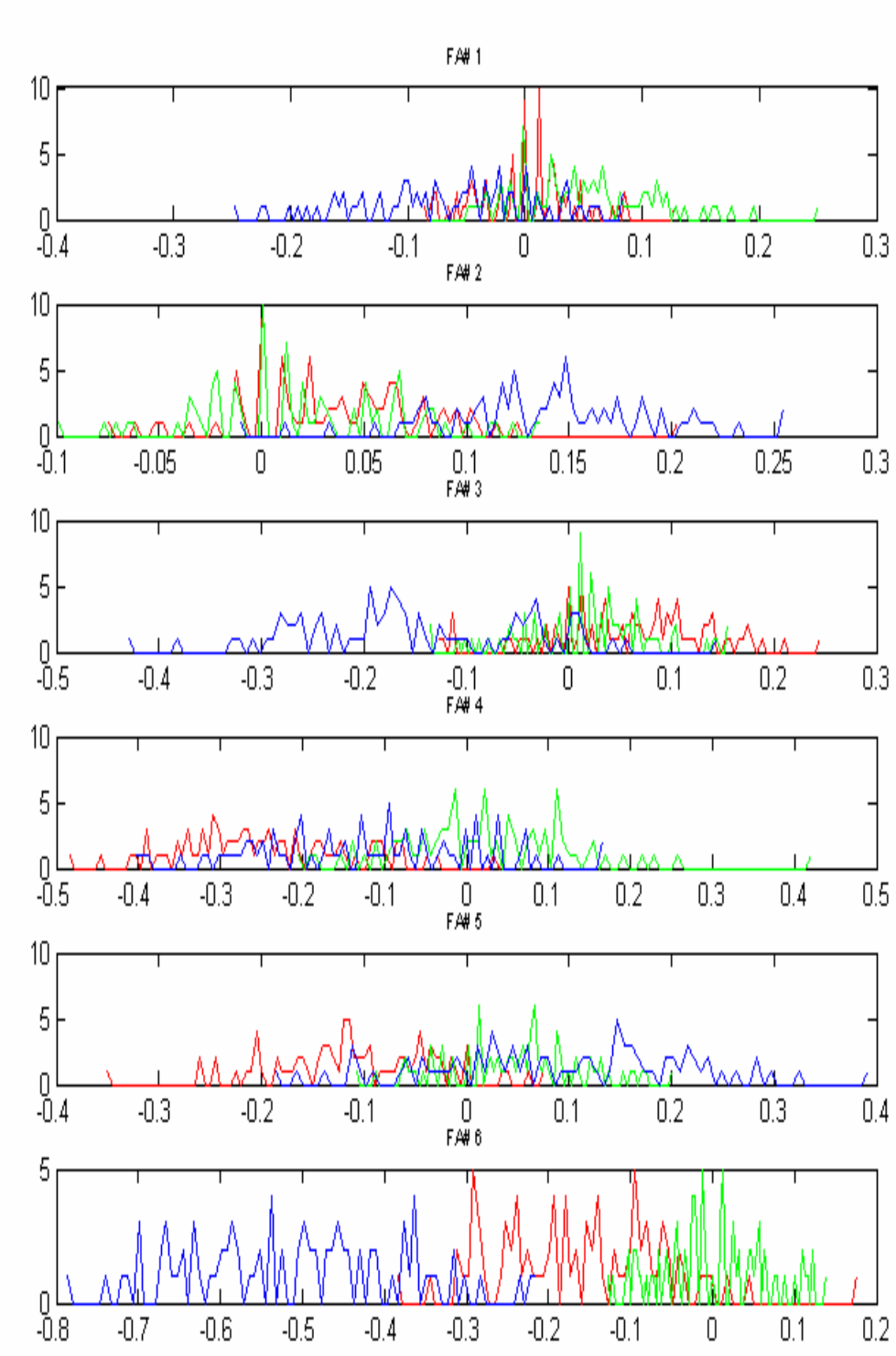


Figure 5-4 Histogram of Facial Actions

(Blue: Scream, Green: Anger, Red: Smile)

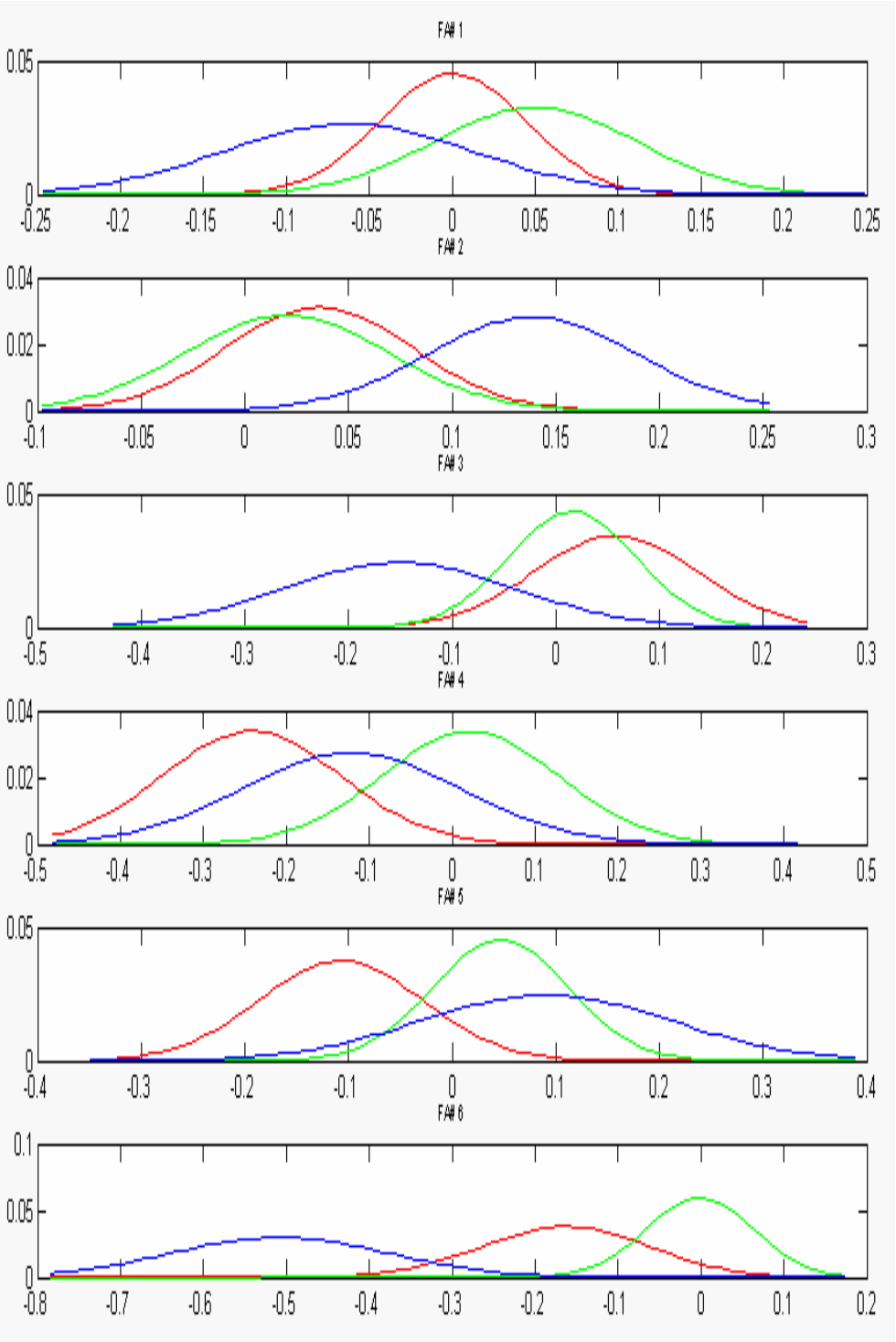


Figure 5-5 Calculated Gaussian Density Functions for Facial Actions

(Blue: Scream, Green: Anger, Red: Smile)

Facial actions are calculated using features extracted with FAFE. Bayes classification is used to classify test images into three expressions. Used performance measure is the ratio of correct classification to image count and named as recognition in this study. Recognition rate on Aleix Face Database is obtained as 78.75%. Confusion matrix is given in Table 5.3.

Table 5-3 Confusion matrix of Bayes Classifier Using FAFE Features
(RR: Recognition Rate)

Smile	**Anger**	**Scream**	**I/0**
68	6	26	**Smile**
7	73	6	**Anger**
5	1	48	**Scream**
85	91	60	***RR(%)***

Results given above contain false classifications due to both feature extraction accuracy and classifier performance. Effect of FAFE accuracy in classifier recognition rate is removed using hand marked feature locations. Recognition rate of Bayes classifier for hand marked data is obtained as 96.25% and confusion matrix is given bin Table 5.4.

Table 5-4 Confusion matrix of Bayes Classifier Using Hand Marked Features

Smile	**Anger**	**Scream**	**I/0**
77	3	3	**Smile**
3	77	0	**Anger**
0	0	77	**Scream**
96	96	96	***RR(%)***

Recognition rate of other classifiers (Nearest Mean and K-Nearest Neighbor) using different distance measures on classified facial actions are obtained. Results are given in Tables 5.5 and 5.6.

Table 5-5 Recognition Rate of Different Classifiers Using FAFE Features (%)

CLASSIFIER	MAHALANOBIS	COSINE	EUCLIDIAN
NM	74.58	62.08	81.25
KNN(K=1)	65.8	55.0	75
KNN(K=2)	67	56	80.83
KNN(K=4)	65.4	52	79.6
KNN(K=8)	63.75	47.08	79.58

Table 5-6 Recognition Rate of Different Classifiers for Hand Marked Features (%)

CLASSIFIER	MAHALANOBIS	COSINE	EUCLIDIAN
NM	92.5	76.25	92.92
KNN(K=1)	76.25	59.58	91.67
KNN(K=2)	78.75	55.00	90.83
KNN(K=4)	80.42	58.33	92.5
KNN(K=8)	78.75	55.83	93.33

Bayes classifier gave the highest recognition rate on hand marked data set (96.25%) and nearest neighbor classifiers performed best (81.25%) when facial actions are extracted using FAFE.

Euclidian distance gave the highest recognition rate for each classifier. Cosine and Mahalanobis distance measures are scale invariant measures and are used to give equal importance to different facial actions with different scales. In Euclidian distance measure, entries with higher scale are more effective for distance calculations. When used for expression classification this distance measure increases the importance of larger scale facial actions. Larger scale facial actions can be thought as less noise sensitive features and may give stronger clues about expression. Higher recognition rate obtained using Euclidian distance measure has a consistency with this idea.

5.5.Holistic Expression Classifier Performance

Holistic expression classification technique implemented in this work uses a set of Gabor filtered images (Gabor Jets) tuned to different scale and orientations for classification.

Two studies [20][18] using a similar approach have been reported. First a description of results obtained in these studies will be presented in section 5.5.1. Performance of implemented holistic expression classifier will be described in section 5.5.2

5.5.1.Gabor Filter Based Expression Classifiers Proposed in Literature

I.Bucio and C.Kotropoulos [20] used Gabor jets constructed using 3 frequencies and 4 orientations and tested their classifier performance using JAFFE database. They manually extracted interior region (80x60) of each image and Gabor jets are calculated using these extracted images. Calculated jets are downsampled by 4 resulting in a jet size of 3600. The same procedure repeated for a higher frequency Gabor jet generation. They used only the magnitude of extracted jets. Leave one out strategy is used for training however there is no clear information about whether they used images of same person for training or not.

Recognition rates obtained in their study using different distance measures for nearest neighbor classifier and SVM classifier are shown in Table 5.7 (Classifier performances using independent component analysis (ICA) are also reported in their study however recognition rate obtained using ICA is below the recognition rate they obtained using Gabor wavelets and they are not written in this table)

Table 5-7 Expression Recognition Rates (%) Reported in [20]

(BRR: Best Recognition Rate, WRR: Worst Recognition Rate, ARR: Average Recognition Rate, HF: High Frequency Gabor Jets, LF: Low Frequency Gabor Jets)

	BRR	Expression	WRR	Expression	ARR
HF+SVM	94.70	DISGUST	87.18	SADNESS	90.34
LF+SVM	93.79	DISGUST	84.52	NEUTRAL	89.44
HF+CDM	90.13	DISGUST	79.15	NEUTRAL	84.17
LF+CDM	88.64	DISGUST	77.41	SADNESS	83.70
HF+EDM	88.74	DISGUST	78.27	NEUTRAL	83.54
LF+EDM	88.19	DISGUST	78.10	NEUTRAL	83.37

Another study using Gabor jets for facial expression analysis is reported by J.Lyons [18]. Gabor jets are calculated using five frequencies and six orientations. PCA is

used to decrease the dimensionality of Gabor jets and nearest mean classifier is used for expression analysis. They reported a classification rate of 75% on JAFFE database.

5.5.2.Holistic Classifier Performance Obtained In This Study

In this study Gabor jets constructed using 6 orientations and 6 scales are used for expression analysis. A sample image and corresponding Gabor jets are shown in Figures 4-9 to 4-11. Performance of K-NN, NM classifiers are tested using cosine and Euclidian distance measures. PCA is used for decreasing the dimensionality of Gabor jets and same classifiers are tested using this low dimensional data. Performance of these techniques is tested using JAFFE and AF databases.

Performance of the classifier on JAFFE database is tested using two training strategies. As a first strategy JAFFE database is divided to 10 groups (according to subjects in the database) and classification of each picture is done using training images from others groups. In this way classifier has no priori information about the person in tested image. Obtained performance using different distance measures and classifiers using leave one out strategy are as shown in Table 5.8.

Table 5-8 Obtained Expression Recognition Rates in JAFFE Database)
(BRR: Best Recognition Rate, WRR: Worst Recognition Rate, ARR: Average Recognition Rate %)

			BRR	EXPRESSION	WRR	EXPRESSION	ARR
OTHER IMAGES OF THE SAME PERSON NOT INCLUDED IN TRAINING	NM	CDM	83	NEUTRAL	31	FEAR	59,62
		EDM	83	NEUTRAL	10	FEAR	59,15
		PCA+CDM	97	SURPRIZE	17	FEAR	61,50
		PCA+EDM	83	NEUTRAL	34	FEAR	60,10
	KNN (K=8)	CDM	97	NEUTRAL	13	FEAR	49,70
		EDM	90	NEUTRAL	24	FEAR	49,29
		PCA+CDM	71	NEUTRAL	23	FEAR	47,88
		PCA+EDM	77	NEUTRAL	27	FEAR	50,07
OTHER IMAGES OF THE SAME PERSON INCLUDED IN TRAINING	NM	PCA+CDM	100	NEUTRAL	84	HAPPY	94,80
		PCA+EDM	100	NEUTRAL	81	HAPPY	94.3
	KNN (K=2)	PCA+CDM	100	NEUTRAL	71	HAPPY	84,00
		PCA+EDM	100	NEUTRAL	71	HAPPY	84,00

When images containing same subject with the test image are included in training set recognition rate significantly increases. Nearest mean and KNN classifiers select members of training set which are closest to test image. When pictures of a subject are both included in test and training sets, classifiers almost every time select those members of training sets which belong to the subject in test image. As a result inclusion of a priori knowledge about a subject significantly increases recognition performance. This fact can clearly be seen from results obtained.

When no priori information about subjects in test images is present nearest mean classifier using both PCA and Euclidian distance measure gave the highest performance in the sense of largest WRR. Confusion matrix for this classifier is given in Table 5.9.

Table 5-9 Confusion Matrix of JAFFE Database

(RR: Recognition Rate %)

ANGER	**DISGUST**	**FEAR**	**HAPPY**	**NEUTRAL**	**SAD**	**SURPRIZE**	
21	7	5	1	2	0	0	**ANGER**
5	15	1	0	0	7	0	**DISGUST**
0	2	11	0	0	4	3	**FEAR**
0	0	1	22	0	1	1	**HAPPY**
1	1	3	7	25	1	6	**NEUTRAL**
3	4	6	0	0	15	0	**SAD**
0	0	5	1	3	3	20	**SURPRIZE**
70	***52***	***34***	***71***	***83***	***48***	***67***	***RR***

Recognition performance of holistic classifier is also tested using AF database. Number of images in AF database is large enough to separate it into strict test and training sets. Performance of holistic classifier is evaluated using the same training and test sets that are used for feature based classifier. Using the results obtained on JAFFE database a nearest mean classifier both using PCA and Euclidian distance measure is selected for AF database. The effect of principal component count on recognition rate is analyzed and a plot showing their relation is given in Figure 5-6. The confusion matrix obtained using 30 principal components is given in table 5.10.

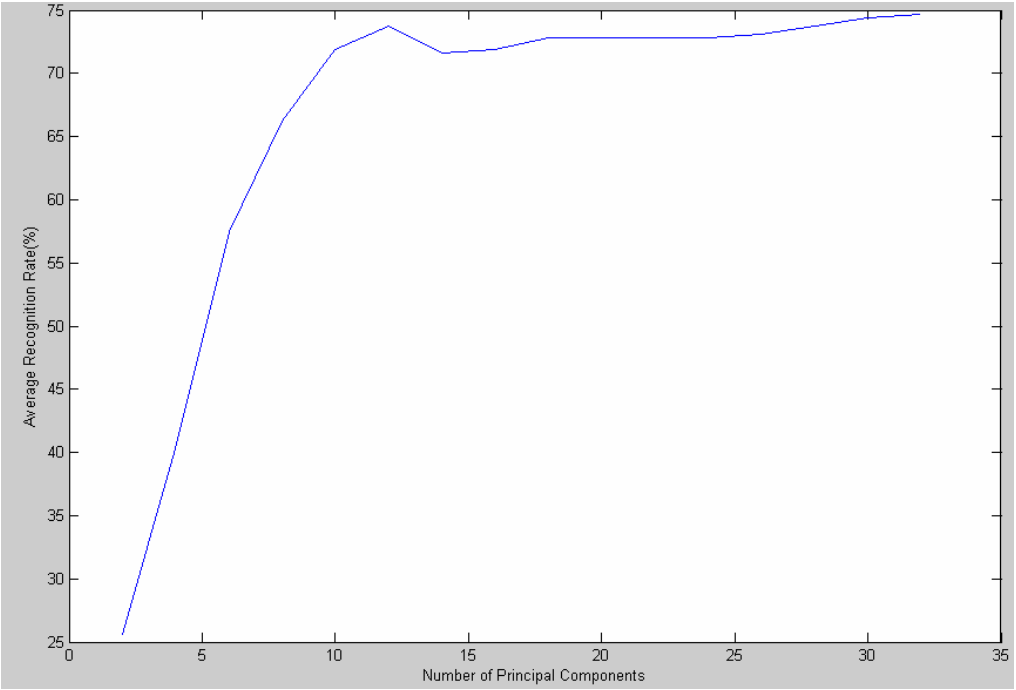


Figure 5-6 Change of Recognition Rate with Changing Principal Component Count for AF database

Table 5-10 Confusion Matrix of AF Database and Holistic Classifier (RR: Recognition Rate %)

NEUTRAL	**SMILE**	**ANGER**	**SCREAM**	***I/O***
53	3	27	0	**SCREAM**
7	67	2	8	**ANGER**
19	6	48	1	**SMILE**
1	4	3	71	**NEUTRAL**
66	***84***	***60***	***89***	***RR***

CHAPTER 6

CONCLUSION

In this thesis a still image facial expression recognition technique has been developed. Proposed technique can be thought as a two step approach. First facial features are extracted by using a fully automatic facial feature extractor (FAFE) and later changes in facial distances are analyzed using a Bayes classifier. Feature extraction starts with ellipse fitting on skin extracted binary image. Orientation and size of extracted ellipse is used for defining search regions for upper and lower facial regions. Once locations of these regions are extracted templates for upper and lower facial features are placed on these regions. Template parameters are recovered using PCA. As another possible usage of FAFE a facial feature tracking system using KLT tracker and FAFE feature extractor is proposed. Features extracted from target and neutral expression images are used for expression classification. Performances of different classifiers are compared.

As an alternative to proposed feature based method a holistic method using Gabor wavelet filters is also implemented. Gabor jets are constructed using multi orientation and multi scale filters. PCA is used for decreasing the dimensionality of calculated jets. Performances of different classifiers using Gabor jets are compared.

Tests on AF database showed that proposed feature extractor can robustly detect facial features on still images. Facial features of images containing large facial occlusion (facial hair, glasses, hair) are successfully extracted. The largest average feature extraction error (corresponding to rightmost of right eyebrow) for proposed technique is observed to be 10.1 pixels on 576x768 images. This localization performance can be accepted to be satisfactory.

Average recognition rate on AF database obtained using Bayes expression classifier and FAFE is found to be 78.75% with a best recognition rate of 91% (anger) and worst rate of 60% (scream). The low performance obtained for scream expression is thought to be caused by large feature extraction error for this expression which can be described by large deformation around eyes and mouth regions. Pure classifier performance on hand marked feature locations is found to be nearly perfect with a recognition rate of 96.25%.

Performance of an available [20] expression classifier technique based on Gabor filters is tested on JAFFE and AF databases. Obtained recognition rates for JAFFE database does not match with results obtained in [20]. Although average recognition rate obtained using cosine distance measure and nearest mean classifier is reported to be 84.17%, training strategy used for this classifier is not clearly presented. Average recognition rate is obtained as 59.62% when no priori information about test subject is given to the classifier and an average recognition rate of 94.8% is obtained when training set includes other images belonging to test subject. In order to make a comparison between Gabor wavelet based technique and proposed technique recognition rates for Gabor wavelet classifier are tested on AF database and an average recognition rate of 75% is obtained.

Proposed classifier requires the availability of neutral expression and this is a disadvantage when compared with Gabor classifier. However Gabor classifier requires manual extraction of inner face region which is not a requirement for proposed classifier. Results on AF database showed that proposed expression classifier is superior in performance to Gabor classifier.

REFERENCES

[1] Ashish Kapoor, Rosalind W. Picard, "Real-Time, Fully Automatic Upper Facial Feature Tracking", Fifth IEEE International Conference on Automatic Face and Gesture Recognition, 2002.

[2]Mathew Turk, Alex Pentland, “Face recognition using eigenfaces”, Media Laboratory, MIT Press, 1991.

[3] Shinjiro Kawato, Jun Ohya, "Real-Time Detection of Nodding and Head-Shaking by Directly Detecting and Tracking the Between-Eyes", Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000.

[4] S. Spors, R. Rabenstein "A real-time face tracker for color video", IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001.

[5] Henry A. Rowley, Shumeet Baluja, Takeo Kanade, “Rotation Invariant Neural Network-Based Face Detection”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA, June 23-25, 1998.

[6] Eli Saber, A. Murat Tekalp “Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions”, Pattern Recognition Letters 19(8) : 669-680 (1998)

[7] Ming-Hsuan Yang, David J. Kriegman, Narendra Ahuja, "Detecting Faces in Images: A Survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24,no.1, pp.34-58, Jan, 2002.

[8]L.Sirovich and M.Kirby “Low-dimensional procedure for the characterization of human faces.” J. Optical Soc. Am, 4:519-524, 1987

[9] Ashok Samal and Mark Propp, “Artificial Neural Network Architectures for Human Face Detection, in Intelligent Engineering Systems Through Artificial Neural Networks”, Volume 2, pp 535-540, November 1992.

[10] T.N.Bhaskar, Foo Tun Keat, S.Ranganath, Y.V Venkatesh, “Blink detection and Eye Tracking for Eye Localization” TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region Volume 2, p.821-824 Vol.2 Oct 2003

[11]A.Haro, M.Flickner, I.Esaa. “Detecting and tracking eyes by using their physiological properties, dynamics and appearance.” In IEEE CVPR, 2000.

[12] Shaogang Gong, Stephen McKenna and Alexandra Psarrou “Dynamic Vision from Images to Face Recognition”, Imperial College Press, 2000

[13] J.G. Daugman, “Uncertainty relation for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical filters”. J. Opt. Soc. Am., 2(7) 1160-1169, 1985.

[14] J.P. Jones and L.A. Palmer. “An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex.” Journal of Neurophysiology, 58(6) 1233-1258, December 1987.

[15] Field D.J. “Relations between the statistics of natural images and the response properties of cortical cells” Journal of The Optical Society of America 4(12): 2379-2394, 1987

[16] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Hurtz, and W. Konen. “Distortion invariant object recognition in the dynamic link architectures.” IEEE Trans. on Computers, 42(3):300-311, March 1993.

[17] Z. Zhang, M.Lyons, M.Schuster, S.Akamatsu. “Comparison Between Geometry-Based And Gabor-Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron.” International Workshop On Automatic Face And Gesture Recognition, Pages 454–459, 1998.

[18] M.J. Lyons, J. Budynek, S. Akamatsu, "Automatic classification of single facial images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 12, pp. 1357-1362, December 1999.

[19] Xiang Cao, Baining Guo. “Real-time tracking and imitation of facial expression.” Proceedings of ICIG 2002, SPIE International Conference on Image and Graphics. p. 910-918, 2002.

[20] Ioan Buciu, Constantine Kotropoulos, Ioannis Pitas ”ICA and Gabor representation for facial expression recognition”. ICIP (2) 855-858 , 2003

[21] Gökberk, B., Irfanoglu, M.O., Akarun, L., Alpaydin, E. “Optimal Gabor kernel location selection for face recognition.” Proceedings of the IEEE International Conference on Image Processing. (2003).

[22] L. Wiskott. “Phantom faces for face analysis.” In Proc. IEEE Intern. Conf. on Image Processing, ICIP'97, Santa Barbara, p 308-311. IEEE, October 1997.

[23] Haisong Gu Zhu and Qiang Ji, “Information Extraction from Image Sequences of Real-world Facial Expressions”, Machine Vision and Applications, Vo. 16, No. 2, p105-115, 2005.

[24] G. Yang, T.S. Huang, "Human face detection in complex background", Pattern recognition, 27(1) 53, 1994.

[25] M.F.Augusteijn and T.L.Skujca, “Identification of Human Faces through Texture-Based Feature Recognition and Neural Network Technology,” Proc. IEEE Conf. Neural Networks, pp. 392-398, 1993.

[26] H.P. Graf, T. Chen, E. Petajan, and E. Cosatto, “Locating faces and facial parts,” IEEE Proc. International Workshop on Automatic Face and Gesture Recognition, pp. 277-282, June 1995.

[27] Y.H. Kwon and N.da Vitoria Lobo, “Face Detection Using Templates,” Proc. Int’l Conf. Pattern Recognition, pp. 764-767, 1994.

[28] C. Kotropoulos and I. Pitas. “A rule based face detection in frontal views.” In ICASP, volume IV, pages 2537-2540, 1997.

[29] P. Gejguš, J. Plaèek, M. Šperka: “Skin Color Segmentation Method Based on Mixture of Gaussians and its Application in Learning System for Finger Alphabet” CompSysTech'2004, Rousse, Bulgaria, 2004.

[30] S. Spors, R. Rabenstein, N. Strobel “Joint Audio-Video Object Tracking” IEEE International Conference on Image Processing (ICIP), Thessaloniki, Greece, October 2001

[31] E. Osuna, R. Freund, and F. Girosi. “Training support vector machines: an application to face detection.” In Proceedings, CVPR, pages 130--136. IEEE Computer Society Press, 1997.

[32] Ying-li Tian, Takeo Kanade, Jeffrey F. Cohn, "Recognizing Lower Face Action Units for Facial Expression Analysis," p.484, Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000

[33] B.D. Lucas and T. Kanade “An Iterative Image Registration Technique with an Application to Stereo Vision” (DARPA) Proceedings of the 1981 DARPA Image Understanding Workshop, p.121-130, April, 1981.

[34] Carlo Tomasi, Takeo Kanade “Detection and Tracking of Point Features Technical Report “ CMU-CS-91-132, April 1991

[35] Y.L. Tian, K. Kanade and J. F. Cohn, "Multi-state based facial feature tracking and detection", Technical Report, Robotics Institute, Carnegie Mellon University, Aug.1999.

[36] M.Malciu , F.Preteux “Tracking facial features in video sequences using a deformable model-based approach.” In: Proceedings of the SPIE. Volume 4121. (2000) 51—62

[37] Craw, I., H. Ellis, and J.R. Lishman, “Automatic extraction of face-features” Pattern Recognition Letters, volume 5, p. 183-187, 1987

[38] Jun Miao, Baocai Yin, Kongqiao Wang, Lansun Shen, Xuecun Chen ”A hierarchical multiscale and multiangle system for human face detection in a complex background using gravity-center template” Pattern Recognition 32(7): 1237-1248 (1999)

APPENDIX A

Binary Image Enhancement Using Morphological Operators

Morphological operations are used to enhance binary and gray level images. In this study only binary morphological operations are used so discussion will be limited to binary morphology.

Intersection of two binary images I and J can be defined as follows. If a point in both binary images is 1 then the same point in their intersection $I \cap J$ is 1 else it is 0.

$$I \cap J = \{p \mid p \in I \quad and \quad p \in J\} \tag{A.1}$$

A point in the union of two images I and J ($I \cup J$) is 1 if the value of the same point in any image is one.

$$I \cup J = \{p \mid p \in I \quad or \quad p \in J\} \tag{A.2}$$

Translation of a binary image I by p is shifting I by an amount of p. If I_p is the translation result, it can be mathematically represented as follows.

$$I_p = \{a + p \mid a \in I\} \tag{A.3}$$

Dilation is a fattening operation which causes an increase in the thickness of target image in every point containing a 1. Structuring element defines the increase in target images size. It is translated to every point which is a member of the target image I and union of translated and target image is calculated. The union of all resultant images gives the dilated version of I by structuring element J. Mathematically dilation of image I with structuring element J can be written as :

$$I \oplus J = \bigcup_{p \in I} I_p \tag{A.4}$$

An example to dilation operation is given below: (In this example and in the following examples center of structuring elements are shown with a • next to the entry)

$$\begin{bmatrix} 0\bullet & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \oplus \begin{bmatrix} 1\bullet & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0\bullet & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Opposite of dilation is erosion operation. Dilation can be thought as a thinning operation of an image I with structuring element J. Those points which are the member of eroded image correspond to regions where we can fit the structuring element completely into the image I. Mathematically this operation can be written as follows.

$$I \otimes J = \{p \mid J_p \subseteq I\} \tag{A.5}$$

An example to erosion operation is given below:

$$\begin{bmatrix} 0\bullet & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \otimes \begin{bmatrix} 1\bullet & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0\bullet & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Erosion and dilation operations are frequently used for filtering binary images. One frequent usage is noise removal. If the nature of noise is known, noisy image can be enhanced by selecting a correct structuring element and applying erosion and dilation operations. Basic operations of binary morphology can be combined and applied in a complex sequence. Erosion operation followed by a dilation operation is called an opening operation. Opening operation can be used for the removal of small regions which cannot be a part of target image.

Let the minimum size of interested regions be nxm. By using a nxm structuring element, erosion will remove all regions smaller than nxm. Interested regions are effected by erosion and thinning occur. Dilation operation following erosion is used to recovering those lost areas in effected regions.

An example to opening operation can be given as follows.

$$\left(\begin{bmatrix} 1\bullet & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \otimes \begin{bmatrix} 1\bullet & 1 \\ 1 & 1 \end{bmatrix}\right) \oplus \begin{bmatrix} 1\bullet & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0\bullet & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Closing operations can be thought as opposite of opening operations. Let some kind of segmentation is applied on a color image and a binary image corresponding to searched region is obtained. Noise and some other factors may affect segmentation and some small holes on resultant binary image may be formed. If the minimum size of possible holes on this object is known prior to segmentation closing operation which is dilation followed by erosion can be used for the removal of false holes on the binary image. If a structuring element of size close to minimum possible holes size chosen and a dilation operation using this structuring element is applied all false holes on binary image is closed. This operation also fattens other regions on image and this effect can be removed by applying erosion with the same structuring element. An example to closing operation is given below:

$$\left(\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \oplus \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right) \otimes \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

APPENDIX B

Principal Component Analysis (PCA)

A method commonly used to reduce the dimensionality of a data set is Principal Component Analysis (PCA). In this method data is assumed to be within a linear subspace of lower dimensionality.

Directions of significant variance are called the Principal components and they can be found by selecting the eigenvectors of covariance matrix with largest Eigen values. Let a vector set $X = \{x_1, x_2, ..., x_M\}$ be constructed using M vectors of size NxN. Mean μ and covariance matrix $\sum$ of the data set are given as follows:

$$\mu = \sum_{m=1}^{M} x_m \tag{B.1}$$

$$X_n = X - \mu \tag{B.2}$$

$$\sum = \mathrm{X_n X_n^T} \tag{B.3}$$

$$\sum = \frac{1}{M} \sum_{m=1}^{M} [x_m - \mu][x_m - \mu]^T \ . \tag{B.4}$$

N^2 x N^2 symmetric matrix $\sum$ characterizes the scatter of the data set X. $\sum$ has N^2 eigenvalues and N^2 eigen vectors. If we select the largest k eigenvalues $\Lambda = [\lambda_1 \quad \lambda_2 \quad . \quad . \quad . \quad \lambda_k]$ and k corresponding eigenvectors $U = [u_1 u_2 \quad . \quad . \quad . \quad u_k]$ with $\Sigma u_i = \lambda_i u_i$, we get the set of eigenvectors corresponding to the k dominant variance in the dataset.

These eigenvectors are mutually orthogonal and span a k dimensional subspace called Principal subspace. N dimensional input x_i can be linearly transformed into a k dimensional vector α by the following transformation:

$$\alpha = U^T (x - \mu) \tag{B.5}$$

An original vector can be reconstructed from the k dimensional subspace using:

$$\tilde{x} = \sum_{j=1}^{K} (\alpha_j u_j) + \mu \tag{B.6}$$

The main idea of Principal Component Analysis is to decrease the dimensionality of a data set from N^2 to k. One problem with this technique is that the number of eigenvectors and eigenvalues to calculate increases with the square of the image size. If we use Σ directly, calculating all of the eigenvalues and eigenvectors become time consuming. Eigenvalues and eigenvectors of the M by M matrix Ω given by equation (B.7) and (B.8) can be calculated much more easily if M<< N^2.

$$\Omega = \mathrm{X}_\mathrm{n}^\mathrm{T}\mathrm{X}_\mathrm{n} \tag{B.7}$$

$$\Omega = \frac{1}{M}\sum_{m=1}^{M}[x_m - \mu]^T[x_m - \mu] \tag{B.8}$$

Let $\Delta = [\delta_1 \quad \delta_2 \quad . \quad . \quad \delta_M]$ be the eigenvalues and $W = [w_1 \quad w_2 \quad . \quad . \quad . \quad w_M]$ eigenvectors of the matrix Ω. The relation between the eigenvectors and eigenvalues of Ω and Σ can be shown as follows:

$$\Omega w_i = \delta_i w_i \tag{B.9}$$

$$\mathrm{X}_\mathrm{n}^\mathrm{T} X_n w_i = \delta_i w_i \tag{B.10}$$

$$X_n \mathrm{X}_\mathrm{n}^\mathrm{T} X_n w_i = \delta_i X_n w_i \tag{B.11}$$

$$\Sigma X_n w_i = \delta_i X_n w_i \tag{B.12}$$

$$\Sigma u_i = \lambda_i u_i \tag{B.13}$$

$$X_n w_i = u_i \;\&\; \delta_i = \lambda_i \tag{B.14}$$

Σ has N^2 eigenvalues where Ω has M eigenvalues. It can be shown that M eigenvalues of Ω corresponds to the largest M eigenvalues of Σ.

By selecting the eigenvectors corresponding to k largest eigenvalues of Ω and multiplying them with X_n, we can calculate the k eigenvectors of interest. If M<< N^2 calculation time of the eigenvectors corresponding to the greatest variations in data set is reduced significantly using Ω in place of Σ.