T.C.

BAHÇEŞEHİR ÜNİVERSİTESİ

AFFECT RECOGNITION FROM FACIAL EXPRESSIONS FOR HUMAN – COMPUTER INTERACTION

M. S. Thesis

Sezer ULUKAYA

İstanbul, 2011

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ
The Graduate School of Natural and Applied Sciences
Electrical and Electronics Engineering

AFFECT RECOGNITION FROM FACIAL EXPRESSIONS FOR HUMAN – COMPUTER INTERACTION

M.S. Thesis

Sezer ULUKAYA

Supervisor: Assoc. Prof. Çiğdem Eroğlu ERDEM

İstanbul, 2011

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ
The Graduate School of Natural and Applied Sciences
Electrical and Electronics Engineering

Title of Thesis   :   Affect Recognition from Facial Expressions for Human
                                               Computer Interaction

Name/Last Name of the Student   :   Sezer ULUKAYA

Date of Thesis Defense   :   18.06.2011

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

                                             Assoc. Prof. Tunç BOZBURA
                                             Acting Director

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

Examining Committee Members:

Assoc. Prof. Çiğdem EROĞLU ERDEM   :   _____

Asst. Prof. Devrim ÜNAY   :   _____

Asst. Prof. Övgü ÖZTÜRK   :   _____

ACKNOWLEDGEMENTS

18.06.2011   Sezer Ulukaya

ABSTRACT

AFFECT RECOGNITION FROM FACIAL EXPRESSIONS FOR HUMAN –
COMPUTER INTERACTION

Ulukaya, Sezer

Electrical and Electronics Engineering
Supervisor: Assoc. Prof. Çiğdem Eroğlu Erdem

June 2011, 62 pages

This thesis first presents a hybrid method for face detection in color images. The well known Haar feature-based face detector developed by Viola and Jones (VJ), that has been designed for gray-scale images is combined with a skin-color filter, which provides complementary information in color images. The image is first passed through a Haar-feature based face detector, which is adjusted such that it is operating at a point on its receiver operating characteristics (ROC) curve that has a low number of missed faces but a high number of false detections. Then, using the proposed skin color post-filtering method many of these false detections are eliminated easily. We also use a color compensation algorithm to reduce the effects of illumination. Our experimental results on the Bao color face database show that the proposed method is superior to the original VJ algorithm and also to other skin color based pre-filtering methods in the literature in terms of precision.

This thesis also presents a Gaussian Mixture Model (GMM) fitting method for estimating the unknown neutral face shape for frontal facial expression recognition using geometrical features. Subtracting the estimated neutral face, which is related to the identity-specific component of the shape leaves us with the component related to the variations resulting from facial expressions. These facial expression related components are then classified using Support Vector Classifiers (SVC). Experimental results on the Extended Cohn-Kanade (CK+) database show that subtracting the estimated neutral face shape gives better emotion recognition rates as compared to classifying the geometrical facial features directly, when the person-specific neutral face shape is not available. We also experimentally evaluate two different geometric facial feature extraction methods for emotion recognition. The first one is based on coordinates of landmark points (CBF) and the second one is based on distances and angles (DABF) between landmarks. The average emotion recognition rates achieved with the proposed neutral shape estimation method and coordinate based features is 88 percent, which is higher than the baseline results presented in the literature, although we do not use the person-specific neutral shapes, and any appearance based features. If we use person-specific neutral face shapes, the recognition rate increases to 94 percent.

In this thesis, a study on an Internet Movie Database (IMDB) plug-in for cast identification in movies is also presented. While watching a movie, the user clicks on the face of the person he is interested in to acquire information. Afterwards, the system

first tries to detect the frontal faces in the clicked frame, and if it cannot find any, a profile face detector is used. The detected face is then tracked backwards and forwards in the shot and a face sequence is obtained. Face recognition is then performed by matching the face sequence extracted from the movie and the face image sets in the training database, which have been collected from the web. IMDB page links of the closest three actors resulting from the matching process is finally presented to the user. In this study, we addressed the following three interesting problems: matching a face sequence and a set of face images, the effects of automatically collected noisy training images on the face recognition performance, and the effects of utilizing prior information of cast on the recognition performance. Experiments have shown that matching between a face sequence and a set of face images is a challenging problem.

ÖZET

İNSAN – MAKİNE ETKİLEŞİMİ İÇİN YÜZ İFADELERİNDEN DUYGU TANIMA

Ulukaya, Sezer

Elektrik-Elektronik Mühendisliği
Tez Danışmanı: Doç. Dr. Çiğdem Eroğlu Erdem

Haziran 2011, 62 sayfa

Bu tezde ilk olarak renkli imgelerde yüz bulma üzerine hibrid bir yöntem sunulmaktadır. İyi bilinen ve gri tonlu imgelerde çalışan Haar öznitelik tabanlı Viola ve Jones (VJ) yüz bulma algoritmasını, renkli resimlerde daha iyi çalışır hale getirmek için, tamamlayıcı bir bilgi olan insan teninin rengini de kullanan bir yöntem önerilmektedir. İmge önce Haar öznitelik tabanlı yüz bulucudan alıcı işletim karakteristiğinde düşük yüz kaybı oranı ve yüksek yanlış yüz bulma oranı olacak şekilde geçirilir. Daha sonra, yanlış bulunan yüzler önerilen ardıl ten rengi filtresi ile kolayca elenir. Ayrıca, aydınlanma etkilerini azaltmak için renk dengeleme algoritması da kullanılmıştır. Bao renkli imge veritabanı kullanılarak yapılan deneyler sonucunda, önerilen algoritmanın orijinal VJ ve literatürdeki ön-filtreleme algoritmasından kesinlik bakımından üstün olduğu görülmüştür.

Bu tezde ikinci olarak, yüz ifadesi tanıma amacıyla, bilinmeyen nötr yüz ifadesinin kestirimi için, yüzdeki geometrik öznitelik vektörleri kullanılarak Gauss Karışım Modelleri'ne dayalı bir yöntem sunulmaktadır. Kestirilen nötr yüz ifadesi, duygulu ifadeden çıkarılarak kişiye bağımlı bileşenler yerine duygu değişimini veren bileşenler elde edilmektedir. Sonra, bu elde edilen duygu değişimine ait bileşenler, destek vektör sınıflandırıcıları kullanılarak sınıflandırılmıştır. Genişletilmiş Cohn-Kanade veritabanı ile yapılan deneylerde, kişiye özel nötr yüz ifadesinin bilinmediği durumlarda yüz öznitelik vektörlerini tek başına sınıflandırmaktansa, kestirilen nötr yüz ifadesinin duygulu yüz ifadesi şeklinden çıkarıldıktan sonra sınıflandırmanın daha iyi tanıma oranları verdiği gözlenmiştir. Ayrıca, duygu tanıma için iki farklı geometrik yüz öznitelik çıkarma yöntemi deneysel olarak değerlendirilmiştir. Bunlardan ilki, takip noktalarının koordinatları tabanlı, ikincisi de bu koordinatlar arasındaki çeşitli mesafe ve açı bilgisi tabanlı yöntemlerdir. Önerilen kestirilen yüz ifadesini çıkaran ve takip noktası koordinat tabanlı öznitelikleri kullanan yöntem ortalama yüzde 88 duygu tanıma oranı vermektedir. Bu oran literatürdeki temel yöntemlere göre daha yüksektir (kişiye özel nötr yüz ifadesi ve görünüm tabanlı öznitelik vektörleri kullanmamasına rağmen). Kişiye özel nötr yüz ifadesi kullanıldığında ise ortalama duygu tanıma oranı yüzde 94'e ulaşmaktadır.

Bu tezde üçüncü olarak, filmlerdeki oyuncuların otomatik tanınmasını ve izleyicinin ilgili Internet Movie Database (IMDB) web sayfasına yönlendirilmesini sağlayan bir

sistemin ön çalışmasına da yer verilmiştir. Sistemde oyuncuların tanınması için gerekli olan eğitim imgeleri, oyuncuların isimleri kullanılarak yapılan Google görsel arama sonucunda elde edilmiştir. İzleyici, filmde ilgisini çeken bir oyuncu hakkında bilgi edinmek için bu oyuncunun ekrandaki yüzüne tıklar. Bunun ardından, sistem ilk olarak ön, eğer bulamazsa, profil yüz bulma işlemini gerçekleştirir. Bulunan yüz, o sahnede ileri ve geri doğru takip edilip, bir yüz dizisi oluşturulur. Oyuncunun kimliğinin saptanması için önceden webden toplanmış, farklı oyunculara ait eğitim yüz imgesi kümeleri ile filmden çıkartılan yüz dizisi karşılaştırılır. Karşılaştırma sonucunda bulunan en yakın üç kişinin IMDB sayfalarının bağlantıları kullanıcıya sunulur. Bu ön çalışmada, yüz tanımayla ilgili şu üç ilgi çekici nokta incelenmiştir: Yüz imgesi kümesi ile yüz dizilerinin karşılaştırılması, metin tabanlı görsel arama sonucu webden otomatik elde edilen ve gürültü içeren eğitim yüz imgelerinin yüz tanıma başarımı üzerine etkisi, ve son olarak, filmdeki oyuncu listesinden yararlanılması ve bu önsel bilgiyi kullanarak karşılaştırma işleminin daha kısıtlı bir sınıf sayısı için yapılmasının başarıma etkisi. Deneyler sonucunda yüz imgesi kümesi ile yüz dizilerinin karşılaştırılmasının zor bir problem olduğu gözlemlenmiştir.

Anahtar Kelimeler: Adaboost, Haar öznitelik vektörleri, Nötr yüz, Gauss karışım modelleri, Ten rengi bulma, Duygu tanıma, Yüz bulma

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Active Appearance Models	: AAM
Active Shape Models	: ASM
Akaike Information Criterion	: AIC
Anger	: An
C plus plus	: C++
Cohn Kanade	: CK
Contempt	: Co
Coordinate Based Features	: CBF
Coordinate Based Features with Neutral Subtraction	: CBF-NS
Discrete Cosine Transform	: DCT
Disgust	: Di
Distance and Angle Based Features	: DABF
Estimated Neutral Subtraction	: ENS
Expectation Maximization	: EM
Extended Cohn Kanade	: CK+
False Negative	: FN
False Positive	: FP
Fear	: Fe
Gaussian Mixture Models	: GMM
Happy	: Ha
Hidden Markov Model	: HMM
Human Computer Interaction	: HCI
Illumination Compensated	: I
Internet Movie Database	: IMDB
K Nearest Neighbour	: KNN
Machine Perception Toolbox	: MPT
Maja and Michel Initiative	: MMI
Matrix Laboratory	: MATLAB
Neutral Subtraction	: NS
Open Source Computer Vision Library	: OpenCV

Principal Component Analysis   : PCA

Point Distribution Models   : PDM

Receiver Operating Characteristics   : ROC

Red Green Blue   : RGB

Sad   : Sa

Support Vector Classifiers   : SVC

Support Vector Machines   : SVM

Surprise   : Su

True Positive   : TP

University College Dublin   : UCD

Viola and Jones   : VJ

Viola and Jones Equal Error Rate   : VJ- EER

LIST OF SYMBOLS

Absolute value : $|\,|$

And operator : $\&$

Average value of the color channel : $R_{ave}, B_{ave}, G_{ave}$

Binary skin color mask : S_j

Blue : B

Class conditional probability density function : $P(.\,|\,.)$

Component densities : $P(s\,|\,G_k)$

Component density parameters : $p(\chi\,|\,\Phi)$

Euclidean distance between two landmarks : $d(.,.)$

Face shape in the first frame of image sequence n : $s_{n,1}$

Face shape vector extracted from an image with a facial expression : $s_{n,i}$

Green : G

Height of the window : h_j^i

Histogram of nonskin pixels : $H_{nonskin}(\underline{c})$

Histogram of skin pixels : $H_{skin}(\underline{c})$

Landmark constant factor : α

Landmark coordinates of the shape : U

Linear scaling factor : S_R, S_G, S_B

Log likelihood function : $\ln p(\chi\,|\,\Phi)$

Mahalanobis distance : D_k

Maximized value of the log likelihood function : L_m

Mean shape vector : $\overline{x}$

Mean vectors of K Gaussian mixture components : μ_k

Mean vector of the kth Gaussian mixture component at the jth landmark : $\hat{\mu}_k^j$

Mixture components : G_k

Mixture proportions : $P(G_k)$

Multiplication operator	: Π
Multivariate Gaussian	: $\aleph(s \mid \mu_k, \Sigma_k)$
Neutral face shape dataset	: χ
Number of components	: K
Original shape vector	: x
Person-specific part of the face shape	: $\hat{s}_{n,i}$
Probability	: P
Red	: R
Skin color threshold	: μ
Skin colored pixels	: C_j^i
Skin pixel	: $\underline{c}$
Standard gray value	: $R_{std}, G_{std}, B_{std}$
The index of best fitting neutral shape	: $k_{n,i}^*$
The number of parameters in the statistical model	: m
Threshold	: τ
Total number of detected windows	: N
Total number of images	: M
Variable part of the face shape	: $v_{n,i}$
Variation weighting matrix	: t
Vector of the j$^{\text{th}}$ landmark in the i$^{\text{th}}$ frame of the n$^{\text{th}}$ image sequence	: $p_{n,i}$
Width of the window	: w_j^i
Window	: V_j^i
20 distance and angle based features	: $f_1 - f_{20}$
16 by 16 covariance matrix	: $\hat{\Sigma}_k$

1. INTRODUCTION

Automatic recognition of facial expressions is a challenging task that has a wide range of applications and has received a lot of interest in recent years. In human-to-human interaction, facial expressions are an important part of the communication. It is foreseen that the ability to recognize human emotions will also be a part of man-machine interaction and ubiquitous affective computing scenarios in the near future (Vinciarelli, Pantic and Bourlard 2009). Therefore, newly emerging human-computer interaction (HCI) scenarios require the recognition of the affective state (sad, angry, happy etc.) of the user. Recognition of the facial expressions is an interdisciplinary task that causes to arise a liaison engineering in the field of human-computer interaction. Image and video processing is crossed with the psychology discipline.

Emotion and affect have a close relation but are distinct terms. Affect means to give a person a stimuli, and which has some impact on the person. After the affect the affected person gives a natural response, which is called emotion to show the influence to the environment. Since these two terms are in a cause and effect process, they generally could be used instead of each other. In this work, affect and emotion are used interchangably because of this close relation (Zeng *et al.* 2009).

There are two main approaches for describing affect. In the *categorical approach* of affect description, emotions are classified into categories. Paul Ekman (1971), specified a set of six emotions: anger, disgust, fear, happiness, sadness and surprise that are universal and associated with muscular patterns in all cultures. The intensity of the emotion can be exaggerated or softened. In the *dimensional approach* of affect description, the emotional states are explained in a multi-dimensional space with the majority of variability described by two dimensions: arousal and valence. Valence measures polarity of the emotions such as positiveness or negativeness and arousal measures the intensity of the emotion (Robotics 2011).

Emotions can also be classified as spontaneous or posed. In posed emotions the subject is instructed to "act", that is to tell a sentence with the specified emotion. Spontaneous emotions are not acted but occurs naturally in an emotional state. Since posed emotions

are not natural, they are generally agitated or exaggerated. Therefore, they are more easily classified as compared to spontaneous ones.

There are various databases, which are widely used for emotion recognition experiments. Some of them contain both spontaneous and posed emotions like MMI (Maja and Michel Initiative) Facial Expression Database (MMI 2011), some of them come with facial tracking data, action units and emotion labels of posed expressions such as the Cohn-Kanade Extended Database (CK+) (Lucey *et al.* 2010), some of them have audio-visual features such as the eNTERFACE'05 Audio-Visual Emotion Database (eNTERFACE'05 2011), and some of them are composed of only women subjects, such as The Japanese Female Facial Expression (JAFFE) Database (JAFFE 2011). There are also more comprehensive databases such as the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database, which contains motion capture, multimodal, multispeaker emotional recordings (IEMOCAP 2011). In Table 1.1, a big picture of the emotional databases is shown.

Table 1.1 Emotion recognition databases in the literature

Database	Spontaneous / Posed	Number of subjects	Video / Image	Publicly available
CK+	Posed	123	Video	Yes
eNTERFACE	Posed	42	Video	Yes
IEMOCAP	Both	10	Video	Partially
JAFFE	Posed	10	Image	Yes
MMI	Posed	50	Both	Yes

Source : This table is modified from Humaine 2011.

The first step of an emotion recognition task is the detection of the face in a given image or image sequence. The problem of face detection refers to determining whether or not there are any faces in a given image and to estimate the location and size of any face (Yang, Kriegman and Ahuja 2002). Face detection is a trivial task for humans, however it is not very easy for computers due to geometric (scale, pose, rotation, facial expressions, occlusion etc.) and photometric variations. The second step of an emotion recognition task is tracking the detected face in time, which is very important to retrieve

data about the emotion to be recognized. Tracking performance has a crucial impact on the emotion recognition task. The last step is the recognition of the affective state of a person by classifying the emotion related data extracted from the facial image sequence into one of the predefined emotion classes.

1.1 MOTIVATION

In HCI scenarios, affect recognition is a crucial task. There are many application areas of affect recognition systems in image the processing world. One of the major application areas of facial expression recognition is human-robot interaction such as the robot iCub (iCub 2011), which is a testbed for cognitive and artificial intelligence reseach applications.

In security-related human behavioral analysis applications, it is important to evaluate the facial expressions of a person. For example, during an interrogation, it is important to know whether the interviewee tells the truth or not (Ryan *et al.* 2009).

Another important application for driver safety is the automatic detection of the drowsiness or fatigue of a driver from head gestures and facial expressions, with the goal of providing a warning system to prevent accidents (Vural *et al.* 2008).

In health-care, it is important to determine real or deceptive pain. Sometimes patients (e.g children) are unable to determine and describe their pain (Ashraf *et al.* 2009). In such cases, it is vital to identify the pain or no pain situations from affect recognition.

Affect recognition can also be used in smart home applications. It can be interesting to hear relaxing music when one comes home in an angry mood.

In this thesis, we concentrate on and present facial image analysis methods for face detection, facial feature extraction, and emotion classification, which are the basic components needed for the above applications.

1.2 LITERATURE REVIEW

Since facial expression recognition has many applications and newly emerging concepts, a lot of research has been done on this topic. An affect recognition system has many components, which are connected to each other. These components can be listed

as: detection of a face, tracking a face, extraction and tracking of facial features, extraction of information about emotions from facial features, and classification into one of the predefined emotion classes. An overview of an affect recognition system is shown in Figure 1.1.
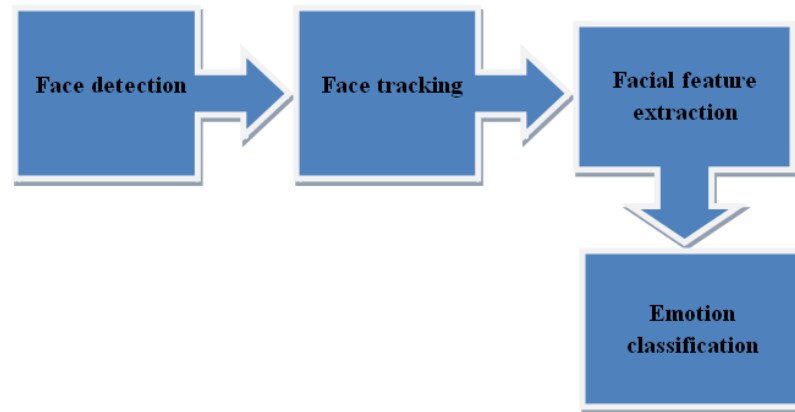
Figure 1.1 : Overview of an affect recognition system

1.2.1 Face Detection and Tracking

First step of a facial expression system is to detect the face in an image. Face detection methods in the literature can be grouped as knowledge-based, feature-based, template-based and appearance-based methods (Zhang and Zhang 2010; Yang, Kriegman and Ahuja 2002; Hjelmas and Low 2001). Knowledge-based methods are a combination of rules that compose a face, while feature-based methods aim to find more distinctive features on the face. Template-based methods aim to find a similar pattern match on the face with the generated template, but appearance-based methods aim to deal with texture structure on the face using pretrained face images (Yang, Kriegman and Ahuja 2002). According to Zhang and Zhang (2010) appearance-based methods are superior to the other methods despite their computational load.

Face detection is a difficult problem because it brings lots of challenges such as non-rigid structure, different illumination conditions, size, orientation, shape, color and texture differences (Yang, Kriegman and Ahuja 2002). According to Yang, Kriegman and Ahuja (2002) pose, presence or absence of components (beards, mustaches and glasses), facial expression, partial or full occlusion, face orientation, lighting conditions

are the basic challenges in face detection. Some of the challenges have been overcome but it still maintains its mystery to be discovered by the researchers.

Face detection is an expensive search problem. In general, a sliding window is scanned through an image at various scales to classify the window as face or non-face. One can say that determining the face as face or non-face is a two-class pattern recognition problem. Therefore, many background windows need to be processed as well as actual face regions. The ratio of the number of non-face windows can be as high as 100000:1. Hence, a well trained classifier is necessary that will produce a low number of false positives. False positive is used here for the regions which are not face but detected as face.

Since face detection is an expensive search problem, a well designed face detection system must be available for real time applications. Face detection based on boosting-based learning algorithms have shown good results (Viola and Jones 2004; Viola and Jones 2001). Viola and Jones (VJ) proposed a frontal face detection system in gray-scale images based on the Adaboost learning algorithm (Viola and Jones 2004). VJ method is suitable for real world applications such as surveillance cameras (Zhang and Zhang 2010). VJ algorithm is available for researchers for the purpose of real time face detection from Open Source Computer Vision Library (OpenCV) tool (OpenCV 2011). Another implementation of similar algorithm is available in the Machine Perception Toolbox (MPT) (Fasel *et al.* 2004).

The number of false detections in the VJ algorithm increases when a high true detection rate is desired. For example, for a database containing 507 faces, there are over 150 false positives to achieve a true detection rate of about 93 percent (Viola and Jones 2004). This false positive rate is too high for some applications, especially in the security domain. The VJ face detector has been reported to fail if the face is tilted beyond about ± 15 degrees in plane, rotated beyond about ± 45 degrees out of plane, towards a profile view. The work of VJ has been extended to handle multi-pose (frontal to profile) faces using skin-color cue (Niyoyita, Tang and Liu 2009) and in-plane rotation using more trained classifiers (Wu *et al.* 2004).

Skin-color is an effective cue for face detection since it is highly invariant to geometric variations of the face such as pose, facial expression and scale and also fast processing is possible (Al Haj *et al.* 2009). Skin-color has been shown to be useful for face detection (Shobana, Yekkala and Eajaz 2007; Hsu, Abdel-Mottaleb and Jain 2002) under varying illumination conditions. There are some approaches in the literature to combine VJ and skin-color to reduce computation time and decrease false detection rate. These approaches are mainly based on pre-filtering method (Shobana, Yekkala and Eajaz 2007) using the University College Dublin (UCD) database (Kim, Ban and Lee 2008) or images from the Internet (Tabatabaie *et al.* 2009). In the pre-filtering method, instead of making an exhaustive search on the whole image, VJ algorithm is applied around probable face regions, where skin-color pixels are highly populated. This results in an improvement in speed and a decrease in the false positive rate. In Chen, Huang and Fu (2008), the authors use a pre-filtering approach to detect candidate face regions and then use a hybrid set of features consisting of Haar-like and Gabor features to train various classifiers for faces in multiple poses but using Gabor features brings extra computational load to the system. In the literature, modified Census Transform is used with skin-color to decrease false detection rate and according to (Wang *et al.* 2008) they have reached a 99 percent true detection rate on the Bao color image database (Frischholz 2011). Methods using neural networks are rotation invariant (Rowley, Baluja and Kanade 1998) but these methods are time consuming for real world applications.

Face tracking is the task of tracking detected faces forward or backward in the scenes in each frame. In general face tracking can be done in two ways. The first one is to use a face detector as a face tracker running at every frame. For example, the VJ face detector can be used as a face tracker running at every frame. The second way is to use a face tracker developed apart from the face detector. In the literature, Active Shape Models (ASM) and Active Appearance Models (AAM) based face trackers are quite popular (Cootes *et al.* 1995; Cootes *et al.* 2001). For the purpose of face tracking, a face shape must be reconstructed to fit to the target face image. ASM is constructed using manually labelled training images (Cootes *et al.* 1995). First, in the image salient points are searched to fit the model and then these points are updated at each frame. This method is also known as Smart Snake since a gradient descent search is done using the face model

to fit (Cootes *et al.* 1995). AAM uses a training phase, too. However, it is better than ASM since it uses both shape and appearance information across the target image (Cootes *et al.* 1995). AAM is slower than ASM, but since it uses all the information available it gives better results than ASM (Cootes *et al.* 1995). The original algorithm is described for gray-scale images but it can be extended to color images (Cootes *et al.* 2001).

AAM based feature extraction is the connection between tracking a detected face and feature extraction and emotion recognition phases. ASM and AAM generation and fitting is described in detail below.

1.2.1.1 ASM generation and fitting

Point distribution models (PDM) are important in modeling of shapes (such as faces), which are easily recognized by humans but not that easily by computers because of their non-rigid nature. PDM uses statistical information about training images to extract knowledge about the mean and the variance of the shape. When describing an object, main characteristics of the shape is used, which are landmarks on the boundary of the object in general (Sonka, Hlavac and Boyle 2008).

In the training part, first all the images are aligned using a transformation. This transformation consists of translation, scaling and rotation. Each image in the training database is co-aligned using this transformation as described in Figure 1.2.

After the algorithm converges, one can say that for any plausible mean shape $\bar{x}$, original shape x can be reconstructed using:

$$x = \bar{x} + U \times t ,$$
(1.1)

where U is the matrix containing eigenvectors of the shape in its columns and t is the variation weighting matrix for each of the eigenvectors.
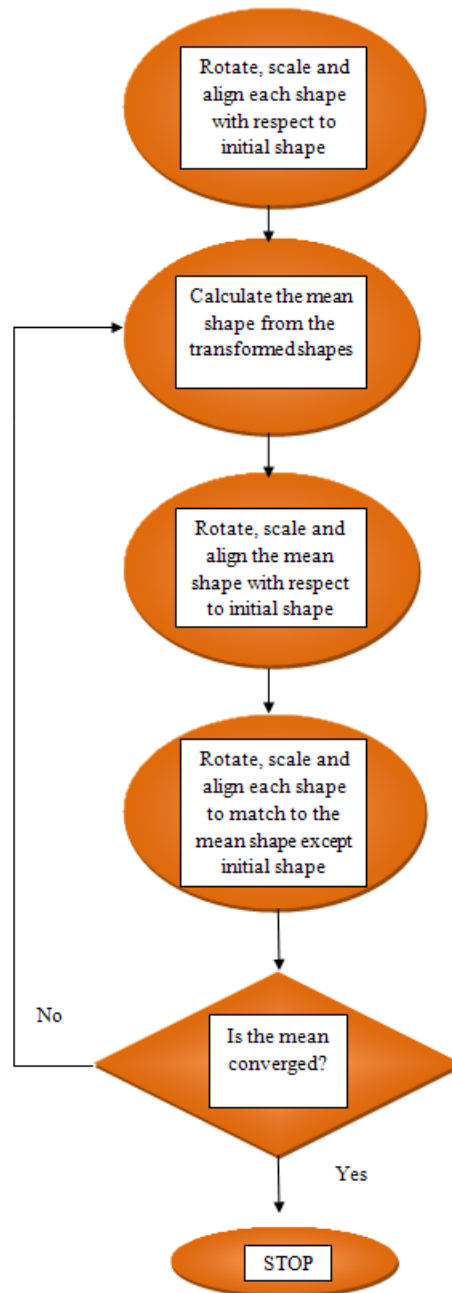
Figure 1.2 : Flowchart of ASM generation

Source : Sonka, Hlavac and Boyle 2008

After the ASM model is generated, it must be fit to the target object. After initialization, landmarks are moved along a search path and the model boundary is fit to the target object boundary as summarized in Figure 1.3 (Sonka, Hlavac and Boyle 2008).
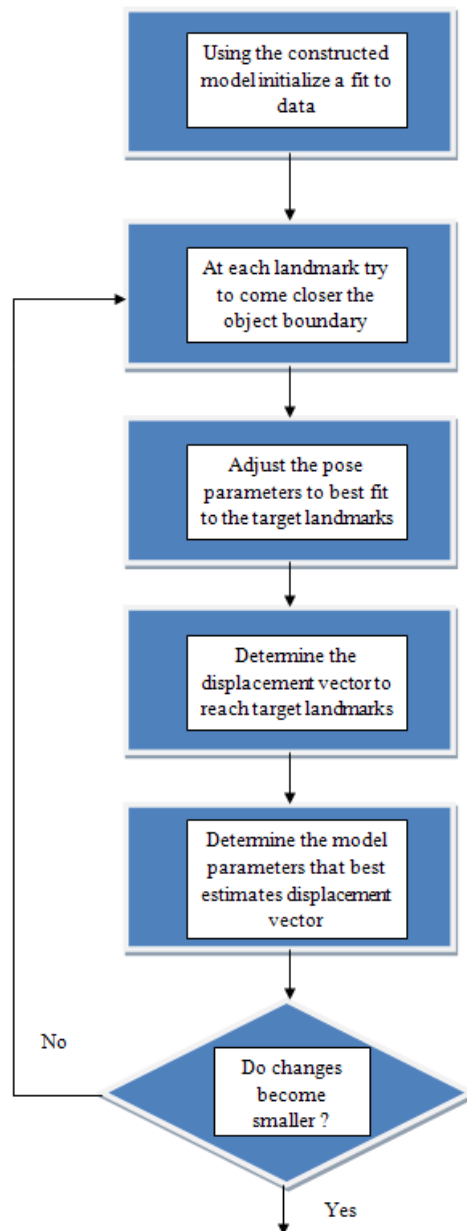
Figure 1.3 : Flowchart of ASM fitting algorithm
Source : Sonka, Hlavac and Boyle 2008

1.2.1.2 AAM generation and matching

In AAM generation the aim is not only to produce a shape model, but also to produce an appearance model to make the model more robust. In ASM, the generated model is solely based on shape but in AAM, the model is described by both shape and appearance

based patch representations (Sonka, Hlavac and Boyle 2008). AAM construction is summarized in Figure 1.4.
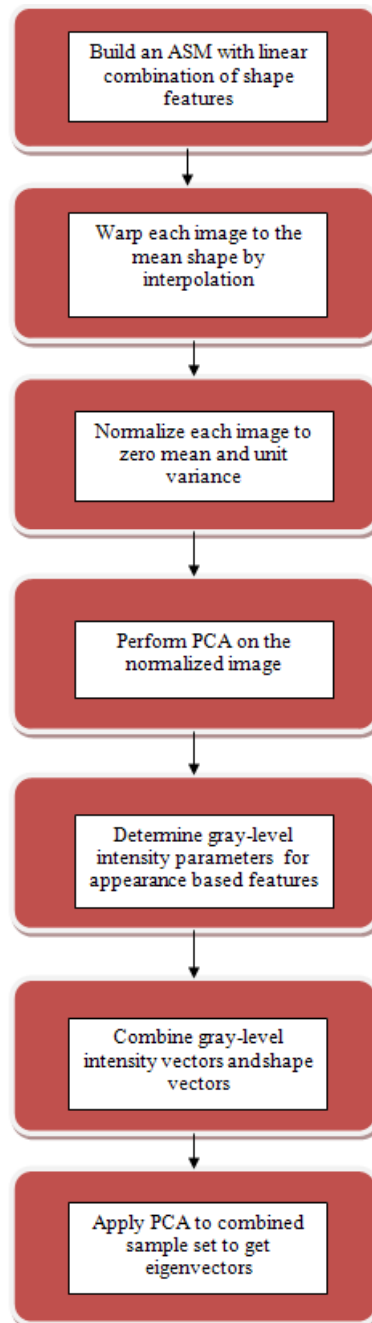
Figure 1.4 : Flowchart of AAM
construction algorithm
Source : Sonka, Hlavac and Boyle 2008

In AAM segmentation global intensity parameters, appearance coefficients and transformation parameters must be optimized (Sonka, Hlavac and Boyle 2008). The algorithm for AAM matching is given in Figure 1.5 in detail.
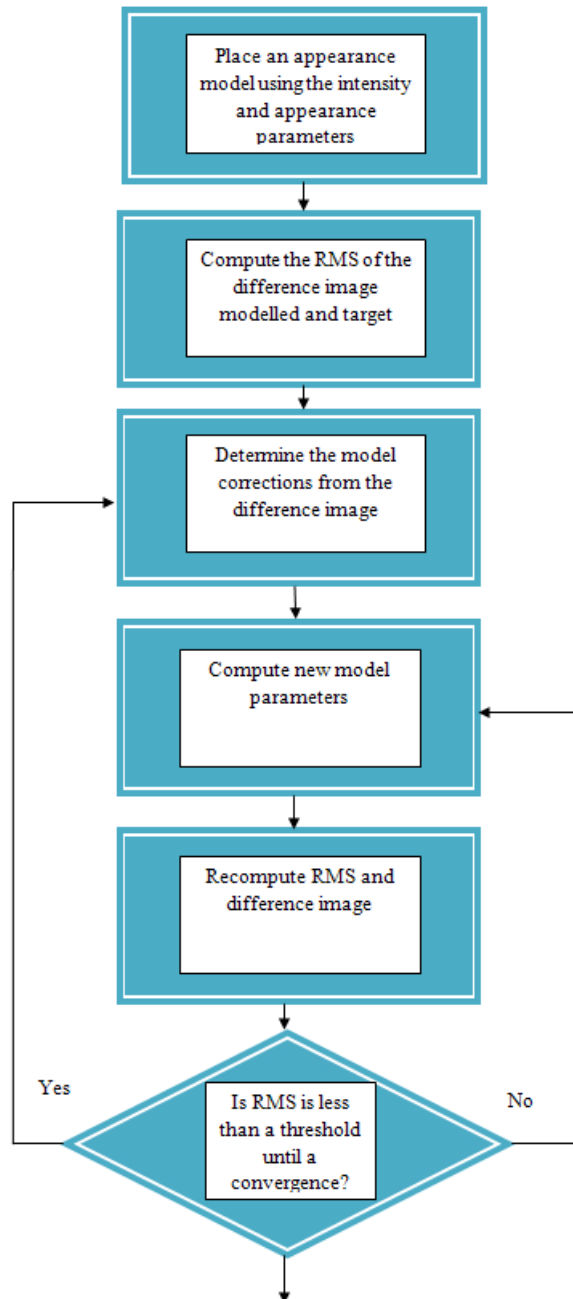
Figure 1.5 : Flowchart of AAM matching algorithm
Source : Sonka, Hlavac and Boyle 2008

It is stated that (Sonka, Hlavac and Boyle 2008), combining AAM and ASM in a hybrid model will increase the model performance.

1.2.2 Facial Feature Extraction and Emotion Classification

Many studies have been published on affect recognition from facial expressions in the last decade, which are summarized in recent survey papers (Gunes and Pantic 2010; Zheng *et al.* 2009; Pantic 2009; Pantic and Rothkrantz 2000). Most of these methods use two dimensional spatio-temporal facial features, which are fed to a pattern recognition algorithm such as Hidden Markov Models (HMM), Support Vector Machines (SVM) and K Nearest Neighbour (kNN) classifiers. Facial features extracted from images or video clips can be broadly categorized as geometrical features and appearance based features. Geometrical features consist of shapes of facial components (eyes, lips etc.) and salient points on the face (nose tip etc.). Appearance based features provide information about the texture of the face as well (natural wrinkles and ceases between the eyes etc.). Geometrical and appearance based features are illustrated in Figure 1.6, where the dots on the left image represent geometrical facial features, which are tracked by an AAM based tracker. One can also represent the wrinkles on the face image of the baby as appearance based features. It is expected that methods that use both geometrical and appearance based features give more accurate results (Zheng *et al.* 2009).

Figure 1.6 : Geometrical features on the left and appearance based features on the right
Source: Left one from Extended Cohn-Kanade (CK+) Database, right one from Google

A major problem in classifying facial expressions is defining the emotion classes. A solution to this has been proposed by Ekman (1971) who specified a set of six emotions:

anger, disgust, fear, happiness, sadness and surprise are universal and associated with muscular patterns in all cultures. Another solution, which has gained popularity recently is to use dimensional and continuous labelling of the affective cues in the valence, activation and dominance coordinates (Gunes and Pantic 2010). A summary of emotion classification approaches in the literature is shown in Table 1.2 below.

Table 1.2 Emotion classification and theorists

Theorists	Emotions
Ekman	Anger, disgust, fear, joy, sadness, surprise
Gray	Rage and terror, anxiety, joy
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
Plutchik	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Tomkins	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise

Source : This table is modified from Ortony and Turner, 1990.

In order to test and compare automatic affect recognition algorithms, databases which are open to researchers are needed. The Cohn-Kanade database has been a very popular one, which consists of facial clips containing the six basic emotions (Kanade, Cohn and Tian 2000) as can be seen in Figure 1.7. The Cohn-Kanade (CK) database has recently been extended to include more subjects, a new facial expression class (contempt) and facial tracking data and is called as the CK+ database (Lucey *et al.* 2010). The face tracking data provided in the CK+ database consists of the locations of 68 points on the face, which are shown on the left image in Figure 1.6 above.

(a) Anger   (b) Disgust   (c) Fear

(d) Happiness   (e) Sadness   (f) Surprise

Figure 1.7 : Examples of six basic facial expressions from the CK+ database

Source: CK+ database(© J. Cohn)

1.3 CONTRIBUTIONS OF THE THESIS

This thesis has four contributions:

1. The first contribution of this thesis is the introduction of a face detection algorithm that combines Haar-feature based and skin color classifiers. The skin color detection is used in a post-filtering framework to decrease the high false positive rate of the well-known Viola and Jones face detector, while keeping a high true detection rate. The windows that are detected as face are verified if the window contains a sufficient number of skin pixels. In order to reduce the effects of illumination, we also use a color compensation method before the skin-color detection step to improve the effectiveness of skin-color detection, which was not present in previous pre-filtering based approaches (Tabatabaie *et al.* 2009; Shobana, Yekkala and Eajaz 2007). The database we used is Bao (Frischholz 2011) color face image database that lacks ground truth data to compare algorithm on the test sets. We prepared the ground truth data for this database and made it available to the researchers (Erdem *et al.* 2011).

2. The second contribution is the experimental evaluation and comparison of two different facial geometric feature computation methods, which we call the coordinate based features (CBF) (Lucey *et al.* 2010) and distance and angle based features (DABF) (Jiao and Pantic 2010). CBF features have been observed

to give higher emotion recognition rates, approximately 94 percent, on the CK+ database (Ulukaya and Erdem 2011).

3. The third contribution is a novel Gaussian Mixture Model (GMM) based method for estimating the neutral face shape for frontal facial expressions using geometrical features. The estimated neutral face, which is related to the identity-specific component of the shape is then subtracted from the currect shape. This provides us the component related to the variations resulting from facial expressions, which are then classified using Support Vector Classifiers (SVC). It is experimentally shown that, subtracting the estimated neutral face shape gives higher affect recognition rates as compared to classifying the geometrical facial features directly, when the person-specific neutral expression is not available (Erdem, Ulukaya, and Erdem 2011).

4. The last contribution of this thesis is the development of a face recognition system using the Internet Movie Database (IMDB) plug-in based on our improved face detector (Ulukaya, Kayim, and Ekenel 2011).

1.4 OUTLINE OF THE THESIS

In Chapter one, the problems of face detection and facial expression recognition are briefly explained and motivated. A literature review on face detection and tracking, emotion recognition and classification is also provided.

In the Chapter two, the proposed method for face detection is presented and explained in detail together with the experimental results.

In Chapter three, the details of the proposed GMM-based estimation of the neutral face shape is given. Experimental results on the CK+ database are given for the neutral face shape estimation for two different geometric feature sets.

In Chapter four, face recognition based IMDB plug-in application is introduced with experimental results.

In Chapter five, a summary and discussion about the thesis is given together with possible future research directions.

2. COMBINING HAAR FEATURES AND SKIN COLOR BASED FACE DETECTORS

This chapter presents a hybrid method for face detection in color images. The well known Haar feature-based face detector developed by Viola and Jones (VJ), that has been designed for gray-scale images is combined with a skin-color filter, which provides complementary information in color images. The image is first passed through a Haar-feature based face detector, which is adjusted such that it is operating at a point on its ROC curve that has a low number of missed faces but a high number of false detections. Then, using the proposed skin color post-filtering method many of these false detections can be eliminated easily. We also use a color compensation algorithm to reduce the effects of lighting. In the following sections, first some background information is provided, which is followed by the details of the proposed face detector.

2.1 BACKGROUND

2.1.1 Adaboost Based Face Detection Using Haar Like Features

VJ (Viola and Jones 2004) have presented a face detection method based on an over complete set of Haar-like features which are calculated in scaled analysis windows. The rectangular Haar-like features are sensitive to edges, bars and other similar structures in the image and they are computed using an efficient method based on the integral image concept. The features used in Viola and Jones (2004) can be seen in Figure 2.1. The integral image concept is faster than any pyramid level detector in the literature and it is similar to the summed area tables in computer graphics (Viola and Jones 2004). Integral image concept deals not with individual image intensities, it is based on the sum of pixel values. As compared to pixel based systems, integral image concept computes the features using only a few operations and this makes it really faster as compared to other algorithms in the literature.

After calculation of a huge number of features for each analysis window, the AdaBoost algorithm is used for combining a small number of these features to form an effective classifier. For example, for an analysis window of size 24x24, there are approximately 160,000 features, far more than the number of pixels. A variant of AdaBoost is used

both to select the best features and to train the final classifier (Freund and Schapire 1997).
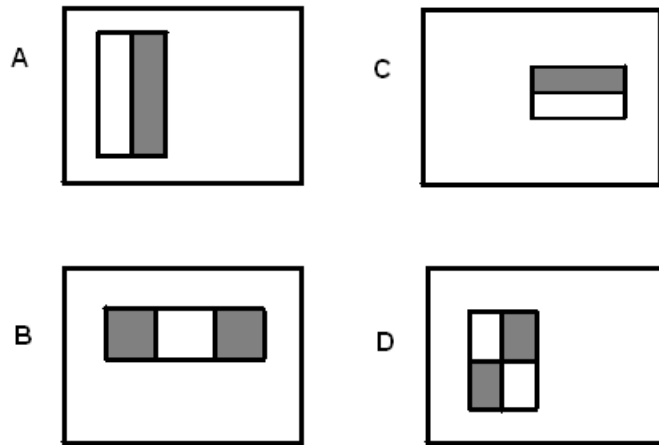
Figure 2.1 : The rectangular feature windows
Source : Viola and Jones 2004

These features are calculated using a few pixel operations as shown in Figure 2.2. In this figure, we want to calculate the area D. The value of the integral image at location 1 is the sum of the pixels in rectangle *A*. The value at location 2 is $A + B$, at location 3 is $A + C$, and at location 4 is $A + B + C + D$. The sum within *D* can be computed as $4 + 1 - (2 + 3)$.
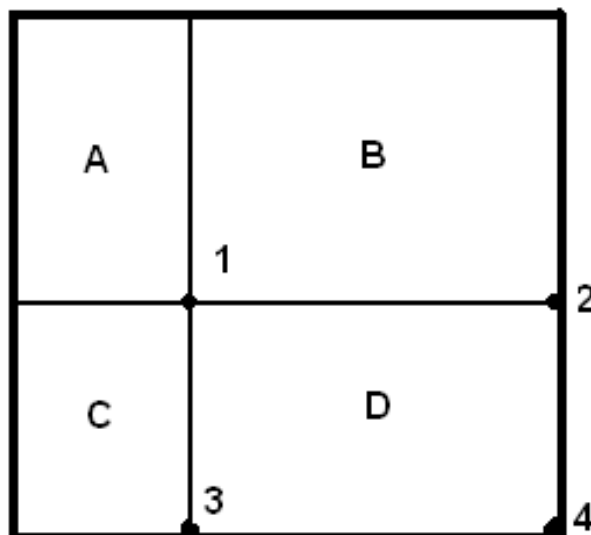
Figure 2.2 : The computation of features
Source : Viola and Jones 2004

VJ face detector has another advantage that is called cascade of classifiers (Viola and Jones 2004). At each step of the cascade, using the features and variations of the features as shown in Figure 2.3, faces are detected as face or non-face in a binary logic concept. For example, in Figure 2.3, feature windows measure the differences of prominent face features; such as darker eye regions and brighter cheeks or darker eye regions and brighter nose bridge to classify face-like regions as face (Viola and Jones 2004). VJ face detector attentional cascade is designed such that filter out 50 percent of the faces while preserving almost all of the faces (Viola and Jones 2004). Once the non-face regions skipped, these regions are not taken into consideration by other stages of the classifiers.
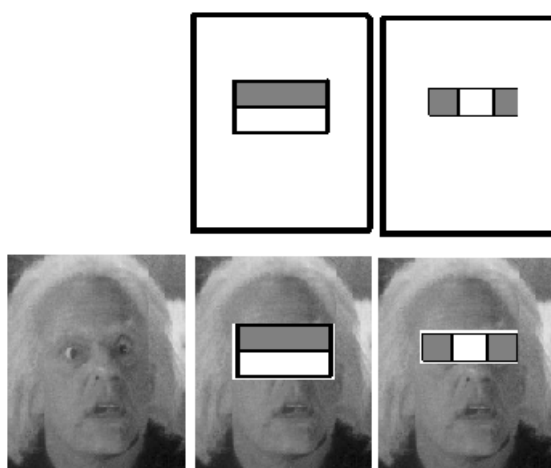
Figure 2.3 : Features on the sample image
Source : Viola and Jones 2004

Failure modes of the VJ face detector depends on the orientation of the features. Since features are horizontal, vertical or diagonal, VJ face detector fails when the rotation is bigger than ± 15 degrees in plane and ± 45 degrees out of plane rotation (Viola and Jones 2004). It is very robust to detect faces when the mouth is occluded and generally fails if the eyes are occluded. There are cases that a person wears sunglasses whose face can be detected by VJ.

2.1.2 Skin Color Detection

The approaches for skin segmentation in the literature have been summarized in several survey papers (Kakumanu *et al.* 2007; Vezhnevets *et al.* 2003). Two methods for skin color detection have been tested in this thesis, which are described below.

2.1.2.1 Bayesian classifier with the histogram technique

The first skin color detection method that is used is based on a Bayesian classifier with histogram technique in Red-Green-Blue (RGB) space. This method has been reported to be superior to other methods in terms of accuracy and computational cost for classifying pixels as skin and non-skin (Phung *et al.* 2005; Jones and Rehg 2002). Using the likelihood ratio method, a pixel with a color vector $\underline{c}$ is classified as a skin pixel if

$$\frac{P(\underline{c}\,|\,skin)}{P(\underline{c}\,|\,nonskin)} \geq \tau, \qquad\qquad\qquad\qquad \textbf{(2.1)}$$

where *P($\underline{c}$|skin)* and *P($\underline{c}$|nonskin)* are the class conditional pdfs of skin and nonskin color distributions, respectively as in equation (2.1). The threshold τ is theoretically proportional to *P(nonskin)/P(skin)*, where the prior probabilities can be estimated from the training set. In practice, the threshold τ is determined empirically, giving a trade-off between the number of false positives and false negatives. A value around *τ = 10.25* gives good results in our experiments. This means that the probability of non-skin pixels is 10 times bigger than the probability of skin pixels.

2.1.2.2 Explicitly defined skin color detector

The second skin color detection method that we test is based on a set of rules on Red, Green, Blue (R, G, B) color components of a pixel. In order to detect the skin colors for a fair complexion under uniform daylight illumination, the following set of rules have been found to be superior to other models under some constraints (Solina *et al.* 2003). A pixel with color components (R, G, B) is detected as skin if the conditions in (2.2) hold. The second line in (2.2) ensures that RGB components must not be close together, which ensures greyness elimination. The third line in (2.2) ensures that R and G components must not be close together, which must be true for fair complexion (Solina *et al.* 2003).

$$R > 95 \,\&\, G > 40 \,\&\, B > 20 \,\&$$
$$\max\{R, G, B\} - \min\{R, G, B\} > 15 \,\& \qquad\qquad \textbf{(2.2)}$$
$$|R - G| > 15 \,\&\, R > G \,\&\, R > B.$$

2.2 HYBRID FACE DETECTOR

Our motivation in this work is to decrease the false positive rate of the VJ face detector (Viola and Jones 2004). The flowchart given in Figure 2.4 shows the steps followed to decrease the false positive rate. Given a color image possibly containing a number of faces, the first step is to apply an illumination compensation algorithm with the goal of reducing the effects of lighting. Then, the image is passed through the VJ detector and a skin-pixel detector. In the next step, the analysis windows that are detected as face by the VJ algorithm are verified by a skin-color based method. Below each step shown in Figure 2.4 will be explained in more detail.
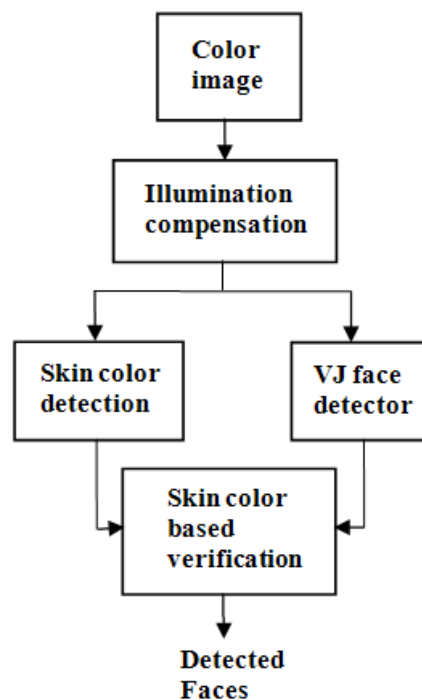
Figure 2.4 : The block diagram of the proposed face detection method

Source: Erdem *et al.* 2011

2.2.1 Illumination Compensation

Illumination compensation is important for eliminating the effects of non-standard illumination for skin color detectors. In this work, we use a color compensation based on the Gray World method (Funt *et al.* 1998), which is fast and simple to implement in

20

RGB color space. This method depends on the assumption that the average surface color in the image is achromatic, which is reflected from the surfaces, corresponds to the color of illumination. The algorithm consists of the following steps:

i) Calculate the averages of each color channel R, G, B for the whole image to get R_{ave}, G_{ave}, and B_{ave}.

ii) Calculate a linear scaling factor for each color component $S_R = R_{std} / R_{ave}$, $S_G = G_{std} / G_{ave}$, $S_B = B_{std} / B_{ave}$, where (R_{std}, G_{std}, B_{std}) denotes the standard gray value.

2.2.2 Skin Color Based Verification

In order to detect the skin-colored pixels in the illumination compensated image, we implemented the two methods described in Section 2.1.2 : a Bayesian classifier with the histogram technique and the explicitly defined skin color detector.

The Bayesian classifier with the histogram technique requires a training step in order to estimate the class conditional pdfs *P(c|skin)* and *P(c|nonskin)*. We estimated these pdfs with the RGB color histograms of skin and nonskin pixels using the Compaq skin database (Jones and Rehg 2002). This database contains wide variability in lighting (indoor/outdoor), background and skin types (white/yellow/brown skins).

In order to estimate the class conditional pdfs of skin and nonskin pixels using histograms, let $\underline{c}$ denote the color vector of a pixel, i.e. $\underline{c} = (r, g, b)$. First we find $H_{skin}(\underline{c})$ and $H_{nonskin}(\underline{c})$, which denote the color histograms of the pixels labeled as skin and nonskin in the training set, respectively. Then the histograms are normalized by dividing by the total number of skin and nonskin pixels. Finally, the skin pixels are detected by applying equation (2.1).

Using either the Bayesian classifier or the explicit detector, binary skin color masks are generated, where skin color pixels are denoted by one and nonskin pixels are denoted by zero. Let S_j denote binary the skin color mask for image j, where M is the total number of images and $j = 1, ..., M$. Let V_j^i denote the ith detected window by the VJ

method for image j, where $i = 1, ..., N$ and N denotes the total number of detected windows claimed as face by VJ method.

Given the detection windows generated by the VJ algorithm and the binary skin pixel masks, the skin color based verification step is carried out as follows:

1. Count the number of skin colored pixels C_j^i in window V_j^i :

$$C_j^i = \sum_{(x,y) \in V_j^i} S_j \qquad\qquad\qquad \textbf{(2.3)}$$

2. Verify window V_j^i as face if:

$$\frac{C_j^i}{w_j^i \times h_j^i} \geq \mu, \qquad\qquad\qquad \textbf{(2.4)}$$

w_j^i and h_j^i denote the width and height of window V_j^i , and μ is a threshold, which is determined experimentally. A value around μ = 0.5 gives good results during the experiments.

2.3 EXPERIMENTAL RESULTS ON FACE DETECTION

In order to evaluate the performance of the proposed method, we used the Bao face database (Frischholz 2011; Wang *et al.* 2008), which consists of color images containing single and multiple frontal and non-frontal faces with a cluttered background. We used the first 100 images of this database containing a total of 859 faces. Since the original Bao database does not contain the ground truth information for the face locations, we first marked ground-truth faces manually in each image by forming a rectangle using the outer corner of the right eye and the left corner of the mouth (see Figure 2.5). If this rectangle is completely within the face detection window generated by VJ algorithm, we define that window as a correct detection. The ground truth data is available from (Ulukaya 2011).
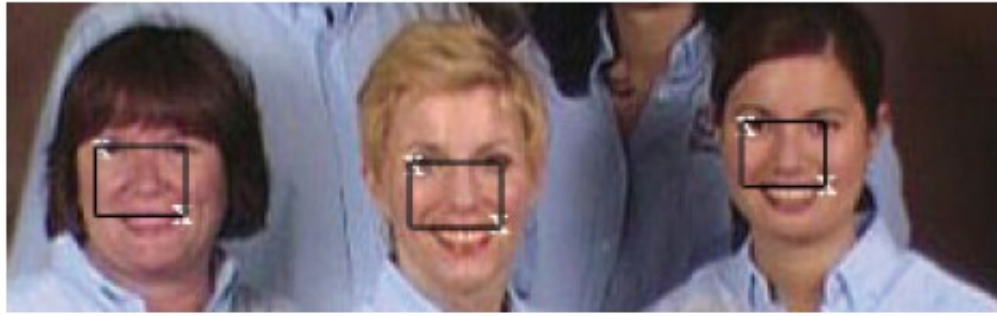
Figure 2.5 :   The manual labeling of the true face locations for the Bao face database.

Source :   A part of image 22 from Bao database is shown.

In Figure 2.5, a rectangle is formed by manually labeling the outer corner of the right eye and the left corner of the mouth, which are shown by the white x signs. We used pre-trained implementation of the VJ face detector in OpenCV library (OpenCV 2011). In order to combine the Haar feature based face detector with the skin-color based post-filtering method, we adjusted the parameters of VJ such that the number of missed faces is as small as possible giving a high correct detection rate. This gives us a high number of false positives, but we expect to eliminate them with the skin color based verification step.

In Table 2.1, we compare the face detection performances of seven methods denoted with the following acronyms:

1. **Bayesian**: Our skin color based post-filtering method using a Bayesian skin classifier.

2. **Bayesian-I**: Our skin color based post-filtering method using a Bayesian skin classifier after illumination compensation.

3. **Explicit**: Our skin color based post-filtering method using an explicit skin classifier.

4. **Explicit-I**: Our skin color based post-filtering method using an explicit skin classifier after illumination compensation.

5. **VJ**: The original Viola and Jones algorithm (Viola and Jones 2004).

6. **VJ-EER**: The original VJ algorithm (Viola and Jones 2004) operating at the equal error rate point of the ROC curve.

7. **Pre-Filter**: The skin-color based pre-filtering method (Tabatabaie *et al.* 2009).

Table 2.1 Performance comparison of seven face detection methods using first 100 images of Bao dataset

Method	TP	FN	FP	Recall	Precision
Bayesian	812	47	39	94.53	95.42
Bayesian-I	811	48	**36**	94.41	**95.75**
Explicit	812	47	46	94.53	94.64
Explicit-I	810	49	40	94.30	95.29
VJ	813	46	237	94.65	77.43
VJ-EER	770	89	83	89.64	90.27
Pre-filter	760	99	77	88.48	90.80

Source : Erdem *et al.* 2011

The acronyms used in the table TP, FP, and FN denote the number of true positives, false positives and false negatives, respectively. The precision and recall are defined as *Precision = TP/(TP +FP)* and *Recall = TP/(TP + FN)*. We can observe from Table 2.1 that the highest precision is achieved by the proposed Bayesian-I method, which is the post filtering method using a Bayesian skin classifier. This precision is much higher than that of the VJ (Viola and Jones 2004) and skin pre-filtering (Tabatabaie *et al.* 2009) methods.

Since the skin color filter is not perfect, it may miss some skin colored pixels. Therefore, the integrity of the face region may not be preserved if the skin color is used as a pre-filter. This causes an increase in the miss (FN) rate of the VJ algorithm which follows the skin color pre-filter, as can be observed in the last row of Table 2.1. However, this effect is not observed if the skin color is used as a post-filter as proposed in this thesis.

If we compare the proposed Bayesian-I and VJ methods given in second and fifth rows of Table 2.1, we can see that the precision increased from 77.43 percent to 95.75 percent, while keeping the recall rate almost the same. We can also observe that using an illumination compensation step is also beneficial in terms of increasing precision. In Figure 2.6 (a), the face detection results of the VJ algorithm are shown for image 26 of the Bao database, where we can see three false detections. In Figure 2.6 (b), the result

after the proposed skin color based post-filtering is shown, where all false positives have been successfully eliminated.

Figure 2.6 (a) : The face detection results of VJ algorithm are shown with squares
Source : 26th image from Bao database (Frischholz 2011)

Figure 2.6 (b) : The face detection results after proposed skin-color post filtering
Source : 26th image from Bao database (Frischholz 2011)

In Figure 2.7 (a), the face detection results of the VJ algorithm are shown for image 39 of the Bao database, where we can see five false detections. In Figure 2.7 (b), the result after the proposed skin color based post-filtering is shown, where all false positives have been successfully eliminated while keeping all faces detected.

**Figure 2.7 (a) : The face detection results of VJ
algorithm are shown with squares**
Source : 39th image from Bao database (Frischholz 2011)

**Figure 2.7 (b) : The face detection results after
proposed skin-color post filtering**
Source : 39th image from Bao database (Frischholz 2011)

In Figure 2.8 (a), the face detection results of the VJ algorithm are shown for image 58 of the Bao database, where we can see four false detections. In Figure 2.8 (b), the result after the proposed skin color based post-filtering is shown, where all false positives have been successfully eliminated while keeping all faces detected.
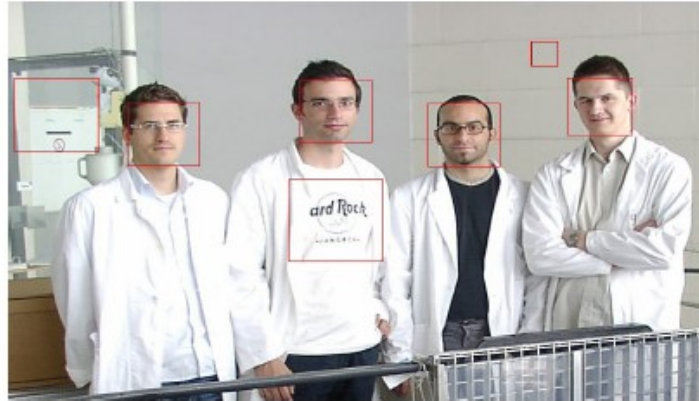
Figure 2.8 (a) : The face detection results of VJ algorithm are shown with squares

Source : 58th image from Bao database
(Frischholz 2011)

Figure 2.8 (b) : The face detection results after proposed skin-color post filtering

Source : 58th image from Bao database
(Frischholz 2011)

The proposed algorithm fails when the number of skin pixels in a face rectangle is larger than the number of non-skin pixels, if the detected face is a false positive. In Figure 2.9 it can be seen that, there are three false detections on the 19th image of the Bao database, but in the below image after the proposed post filtering method one of the false detections can not be eliminated due to the reason explained above. These false detections could be easily eliminated using knowledge based methods such as verifying the existence of a mouth and eyes in the face rectangle.
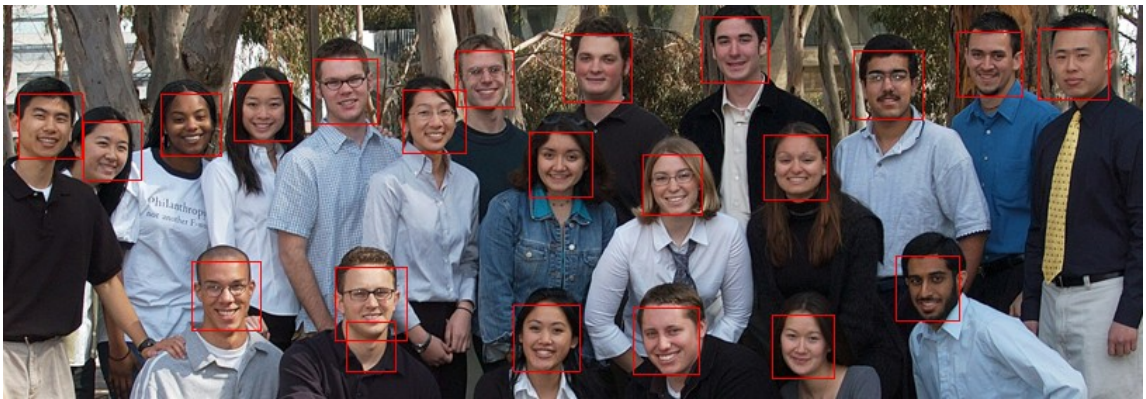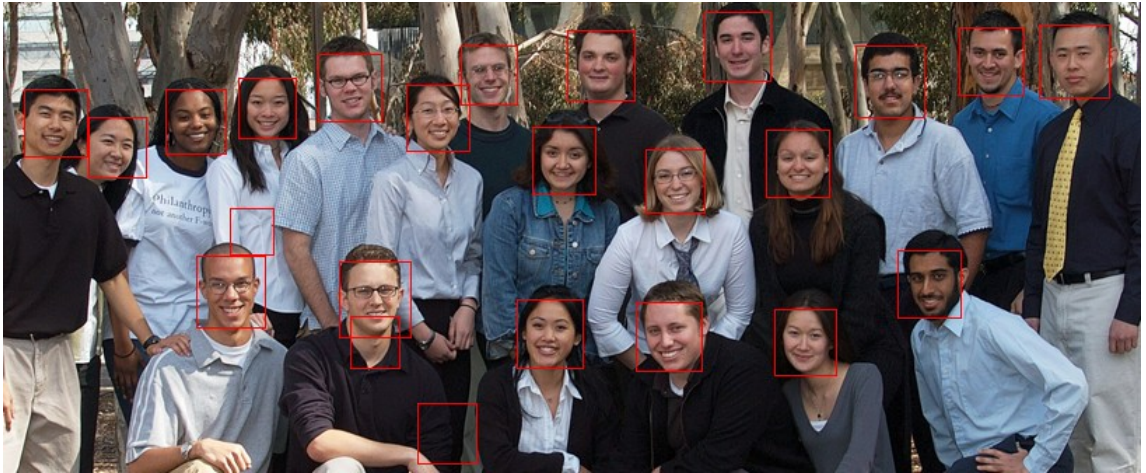
Figure 2.9 : The face detection results after direct VJ (above), and after proposed post filtering method (below). Note that, two of the three false alarms have been eliminated successfully.
Source : 19th image from Bao database (Frischholz 2011)

Table 2.2 Comparison of elapsed time for face and skin detection using first 10 images of Bao dataset

Image number	Skin detection time(ms)	Face detection time(ms)	Number of detected faces
1	2.80	444.28	12
2	4.46	442.85	10
3	2.86	405.04	8
4	1.40	385.37	6
5	3.35	403.14	19
6	2.72	352.35	20
7	2.47	293.47	24
8	8.91	1069.07	13
9	3.86	426.07	10
10	11.79	1186.96	32

The computation time to process the detected face rectangles using the post-filter is very low as compared to the computation time to detect face regions using the VJ algorithm. The elapsed times for several images can be seen in Table 2.2 in miliseconds (ms). In some of the images, elapsed time to detect faces using VJ seem to be high since there are many faces in those images in different scales.

3. FACIAL EXPRESSION RECOGNITION BY NEUTRAL FACE SHAPE ESTIMATION

When we describe a facial expression using locations of a set of points on the face, these geometric locations encode two types of information. The first type information is the identity-specific information, which is constant for that person. The second information is a variable part, which depends on pose and facial expressions. The identity-specific component can be eliminated by subtracting the features obtained from a neutral facial expression of that person from the current frame, which may be the first frame of a video clip (Lucey *et al.* 2010) as in CK+. However, neutral face information of that person may not always be available. In that case, researchers generally average the features of a certain number of images in the video clip, assuming that averaging will resemble a neutral facial expression (Gajsek, Struc, and Mihelic 2010). However, this assumption is not always true, depending on the content of the video clip.

This chapter has two contributions: i) We first present a Gaussian Mixture Model (GMM) based method for estimating the neutral face shape for frontal facial expression recognition using geometrical features, when the person-specific neutral face shape is not available. Subtracting the estimated neutral face, which is related to the identity-specific component of the shape leaves us with the component related to the variations resulting from facial expressions. These facial expression related components are then classified using Support Vector Classifiers. We experimentally show that subtracting the estimated neutral face shape gives better affect recognition rates as compared to classifying the geometrical facial features directly, when the person-specific neutral expression is not available. ii) We also experimentally evaluate two different geometric features, which we call the CBF (Lucey *et al.* 2010) and DABF (Jiao and Pantic 2010). CBF features have been observed to give higher emotion recognition rates on the CK+ database.

This chapter is organized as follows. First, a more detailed description of the used geometrical features is given. Then, the details of the proposed GMM-based estimation of the neutral face shape is provided, which is followed by experimental results.

3.1 GEOMETRICAL FACIAL FEATURES

In this thesis, we utilize the face tracking data provided in the CK+ database to form two types of geometrical facial features as described below. The CK+ database provides the locations of 68 points on the face at each frame, which are tracked using Active Appearance Models (Lucey *et al.* 2010; Cootes *et al.* 2001). Examples of these 68 points can be seen in Figure 3.1 and Figure 3.2.

Figure 3.1 : AAM based tracking of landmark coordinates in CK+ database 1st, 5th, and 33rd frames respectively from left to right

Source : CK+ database(© J. Cohn)

There are 123 subjects and 327 emotion labeled image sequences in the CK+ database. Image sequences start with a neutral (onset) frame and end with a peak frame (apex) of the expression (see Figure 3.1). There are seven emotion categories in the database: anger, disgust, fear, happy, sadness, surprise, and contempt.

Before the geometrical feature vectors are formed for each frame of a video clip, we need to align the face shapes described by the tracked landmark points for all frames in the database to eliminate any rotation, translation and scale effects that may exist between subjects and/or within a video clip.

3.2 ALIGNMENT OF THE FACE SHAPES

Alignment of the face shapes for all frames of the CK+ database is carried out using the landmark points that are affected the least from the facial expressions such as the nose tip and the inner corners of the eyes. The inner corners of the eyes are not affected much from facial expressions and they are robust to track (Jiao and Pantic 2010). First, we

move the nose tip to the origin (point 31 in Figure 3.2). In order to compensate for in-plane head rotations, all the landmarks are rotated such that the line connecting the inner corners of the eye becomes horizontal (i.e., parallel to the x-axis). Another set of points that are expected to be effected from facial expressions the least are the landmarks located at the outer borders of the cheeks (points 1, 2, and 16, 17 in Figure 3.2).

**Figure 3.2 : The 68 landmark points
tracked on the face as given
the CK+ database**
Source : CK+ database(© J. Cohn)

In order to compensate for any scale differences between frames, we scale the landmarks coordinates such that the sum of distances between three point pairs is constant:

$$d(p_{n,i}^{1}, p_{n,i}^{17}) + d(p_{n,i}^{2}, p_{n,i}^{16}) + d(p_{n,i}^{40}, p_{n,i}^{43}) = \alpha, \tag{3.1}$$

where $p_{n,i}^{j} = [x_{n,i}^{j}, y_{n,i}^{j}]$, $k=1,...,M$ denotes the vector representing the jth landmark point in the ith frame of the nth image sequence, and $M=68$. The operator $d(.,.)$ denotes the Euclidean distance between two landmarks. The constant was chosen as $\alpha = 10$ during the experiments. In order to reduce the effects of any tracking errors in landmark

coordinates, we used the sum of distances between three point pairs to normalize for scale differences.

3.3 COORDINATE BASED FEATURES (CBF)

The coordinate based features consist of the x and y coordinates of the M aligned landmarks points in the last (peak) frame of an image sequence (CBF). When the landmarks points of the person-specific neutral facial expression are available (which is the first frame in CK+ database), they can be subtracted from the peak frame, and will be referred to as coordinate based features with neutral subtraction (CBF-NS).

3.4 DISTANCE AND ANGLE BASED FEATURES (DABF)

Another set of geometrical features that we evaluate are derived from the CBF features and they consist of distances and angles between certain landmark points as described below. A total of 20 features ($f_1 - f_{20}$) are obtained from the last frame of an image sequence as follows (Jiao and Pantic 2010) (see Figure 3.3):

Eyebrows: f_1, f_2: angles between the line that connects inner corners of the eyes and the line that connects inner and outer eyebrow (right/left). f_3, f_4: the vertical distance from outer eyebrow to the horizontal line connecting inner and outer corners of the eyes.

Eyes: f_5, f_9: distance between the outer eye (right/left) corner and the upper eyelid. f_6, f_{10}: distance between the inner eye corner and the upper eyelid. f_7, f_{11}: distance between outer eye corner and the lower eyelid. f_8, f_{12}: distance between the inner eye corner and the lower eyelid. f_{13}, f_{14}: vertical distance between the upper eyelid and the lower eyelid.

Mouth: f_{15}, f_{16}: distance between the upper lip and the left/right mouth corner. f_{17}, f_{18}: distance between the lower lip and left/right mouth corner. f_{19}: distance between the left and right mouth corner. f_{20}: vertical distance between the upper and the lower lip.
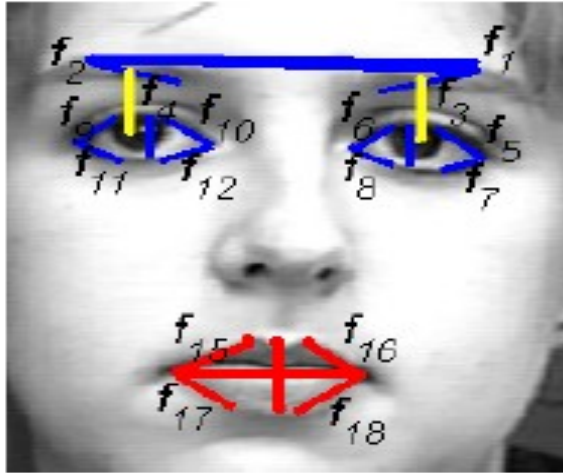
**Figure 3.3 : The 20 distance and angle
based geometrical features
(DABF) of a face**
Source : Features from Jiao and Pantic 2010 and
image from CK+ database (© J. Cohn)

When the person-specific neutral face shape is available, we can subtract the 20 DABF features of the first frame from the peak frame to obtain another set of features that we call as DBAF-NS features, where NS stands for "neutral subtraction".

3.5 ESTIMATION OF THE NEUTRAL FACE SHAPE USING GAUSSIAN MIXTURE MODELS

Neutral face shapes of people in a population are quite different from each other. Some people have long and thin faces while others have round faces. Therefore, we first aim to identify typical face shapes in the population, by fitting a Gaussian Mixture Model to the shape features of neutral faces. We expect that the mean vectors of each Gaussian component to represent a typical face shape cluster.

3.5.1 Fitting a Gaussian Mixture Model to Neutral Face Shapes

The data set of neutral face shapes is constructed from the first frames of all image sequences that are provided in the CK+ database (593 sequences in total) that belong to 123 subjects. Let us represent our neutral shape data set as: $\chi = \{s_{n,1}\}, n = 1, ..., N,$ where $s_{n,1} = [p_{n,1}^1, p_{n,1}^2, ..., p_{n,1}^M],$ represent the face shape in the first frame of image sequence n, based on the normalized coordinates of 68 landmark points. Here the parameters are $M = 68; N = 593.$

34

We want to model the distribution of neutral face shapes using a *mixture of densities* as follows in equation (3.2) (Alpaydın 2010):

$$p(s) = \sum_{k=1}^{K} p(s \mid G_k) P(G_k),$$

(3.2)

where G_k are the *mixture components*, which are also called *clusters*. $p(s \mid G_k)$ are the *component densities* and $P(G_k)$ are the mixture proportions (mixing coefficients). The number of components K is either specified beforehand or can be estimated using Akaike's information criterion as described below. If the component densities are multivariate Gaussian, we have $p(s \mid G_k) \sim \aleph(s \mid \mu_k, \Sigma_k)$ and $\Phi = \{P(G_k), \mu_k, \Sigma_k\}_{k=1}^{K}$ are the parameters that should be estimated from the data set $\chi = \{s_1, ..., s_N\}$. We look for component density parameters that maximize the likelihood of the data set (sample). The likelihood of the sample assuming that the data points are drawn independently from the distribution is as in equation (3.3):

$$p(\chi \mid \Phi) = \prod_{n=1}^{N} p(s_n \mid \Phi)$$

(3.3)

$$= \prod_{n=1}^{N} \left(\sum_{k=1}^{K} P(G_k) \aleph(s_n \mid \mu_k, \Sigma_k) \right),$$

are the log likelihood of the data set is given by as in equation (3.4):

$$\ln p(\chi \mid \Phi) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} P(G_k) \aleph(s_n \mid \mu_k, \Sigma_k) \right).$$

(3.4)

The log likelihood function given in equation (3.4) is maximized using the Expectation-Maximization (EM) algorithm (Bishop 2006).

1. Initialize the means μ_k, covariances Σ_k and the mixing coefficients $P(G_k)$, and evaluate the initial value of the log likelihood.

2. **Expectation step** Evaluate the responsibilities using the current parameter values:

$$\gamma_{kn} = P(G_k \mid s_n) = \frac{P(x_n \mid G_k) \times P(G_k)}{p(s_n)}$$ (3.5)

$$= \frac{P(G_k) \times \aleph(s_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} P(G_j) \times \aleph(s_n \mid \mu_k, \Sigma_k)}$$

3. **Maximization step** Re-estimate the parameters using the current responsibilities:

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{kn} s_n$$ (3.6)

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{kn} (s_n - \mu_k^{new})(s_n - \mu_k^{new})^T$$

$$P(G_k)^{new} = \frac{N_k}{N}$$ (3.7)

where

$$N_k = \sum_{n=1}^{N} \gamma_{kn}$$ (3.8)

4. Evaluate the log likelihood and check for convergence of the parameters or the log likelihood. If convergence criterion is not satisfied return to step 2.

$$\ln p(\chi \mid \Phi) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} P(G_k) \aleph(s_n \mid \mu_k, \Sigma_k) \right)$$ (3.9)

The parameter *K* can be determined experimentally using Akaike's information criterion (Akaike 1974). It is often used to determine an appropriate number of mixture components when the number of components is unspecified. Akaike information criterion (AIC) is the negative log-likelihood for the data with a penalty term for the number of estimated parameters as:

$$AIC = 2m - 2L_m,$$
(3.10)

where *m* is the number of parameters in the statistical model and L_m is the maximized value of the log likelihood function. The GMM fitting process is carried out for a range of *K* values, and the value that maximizes the Akaike Information Criterion (*AIC)* is selected.

After fitting a Gaussian Mixture Model to the data set of neutral face shapes, the mean vectors $\mu_k, k = 1,..., K$ of the *K* Gaussian mixture components will represent the typical neutral face shapes in the population. The covariance matrices Σ_k will represent the variation of the face shapes around the mean shapes.

3.5.2 Estimation of the Neutral Face Shape

Given a shape vector $s_{n,i}$ estimated from image *i* of the nth sequence with a facial expression, we assume that it can be decomposed as follows:

$$s_{n,i} = \hat{s}_{n,i} + v_{n,i},$$
(3.11)

where $\hat{s}_{n,i}$ represents the person-specific part of the shape and $v_{n,i}$ represents the variable part of the shape due to pose and facial expression, which are mostly related to the emotional state of the subject. If the neutral face shape of that person is available, it can be subtracted from $s_{n,i}$, to give the variable part of the shape, which can then be classified.

However, if the person-specific neutral face is not available, it is beneficial in terms of increasing the correct classification rate to estimate the 'best' fitting neutral face shape

and subtract it from $s_{n,i}$. In order to select the best fitting neutral shape among the K face shapes which were estimated using GMM fitting, we use the landmarks that are that are not affected from facial expressions much. The point set selected for this purpose consist of the left and right sides of the cheeks and the inner corners of the eye (see Figure 3.2): $\{p_{n,i}^1, p_{n,i}^2, p_{n,i}^3, p_{n,i}^{15}, p_{n,i}^{16}, p_{n,i}^{17}, p_{n,i}^{40}, p_{n,i}^{43}\}$.

Let us relabel the above points for the ith frame of sequence n as $\{P_{n,i}^j\}$ and let us denote the corresponding points in the mean vector of the kth Gaussian mixture component as $\{\hat{\mu}_k^j\}$, where $j = 1,...,8$ and $k = 1,...,K$. In order to select the best fitting neutral shape we minimize the following Mahalanobis distance:

$$D_k(P_{n,i}^j, \hat{\mu}_k^j) = \sqrt{(P_{n,i}^j - \hat{\mu}_k^j)^T \hat{\Sigma}_k^{-1} (P_{n,i}^j - \hat{\mu}_k^j)} ,$$

(3.12)

where $\hat{\Sigma}_k$ is the 16 × 16 covariance matrix for the x and y coordinates of the landmark points 1, 2, 3, 15, 16, 17, 40, 43, and is formed from the the full covariance matrix Σ_k, which is 136 × 136. The index of the best fitting neutral shape is:

$$k_{n,i}^* = \arg\min_k D_k(P_{n,i}^j, \hat{\mu}_k^j)$$

(3.13)

After the index of the best fitting neutral face is estimated, the mean shape corresponding to that Gaussian mixture is assigned to the person-specific component in (3.6) as:

$$\hat{s}_{n,i} \cong \hat{\mu}_k$$

(3.14)

Hence, the variable part of shape due to the facial expression can be approximated as:

$$s_{n,i} - \hat{\mu}_k \cong v_{n,i}$$

(3.15)

which is classified using a support vector classifier with a second order polynomial kernel (PRTools 2011).

3.6 EXPERIMENTAL RESULTS ON EMOTION RECOGNITION

Experiments are done on the CK+ database (Lucey *et al.* 2010). The Gaussian Mixture Fitting to the neutral face shapes is carried out using the first frames of all sequences for various values $K = 1,\ldots, 11$ and the $K = 6$, which gave the minimum AIC value is selected. During GMM fitting, we used a small non-negative regularization number added to the diagonal of covariance matrices to make them positive-definite.

The mean shapes of the estimated Gaussian mixtures for $K = 6$ is shown in Figure 3.4. Each mean vector is shown with a different marker. We can observe that the estimated mean shape vectors reflect the person-specific variations of the face shape in the population. In Figure 3.5 (a), a happy facial expression of subject 106 in the CK+ database is shown. The CBF features after alignment is given in Figure 3.5 (b), and the estimated neutral face shape is shown in Figure 3.5 (c). For comparison purposes, we draw the CBF features and its estimated neutral face shape together in Figure 3.6 (a). In Figure 3.6 (b), the worst fitting neutral shape is also shown for comparison purposes. We can see that the best neutral face shape (shown blue + signs) follows the person specific characteristics of the face better than the worst fitting neutral shape (shown with gray diamonds), especially if we observe the landmarks around the inner corners of the eyes and the sides of the face. Hence, we can say that the proposed algorithm is successful in estimating a reasonable neutral face shape based on the GMM of the population.
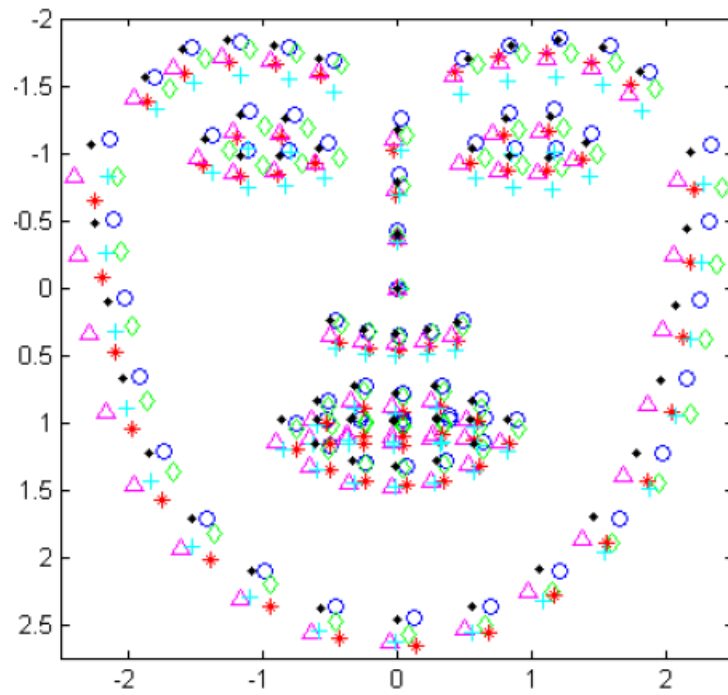
Figure 3.4 : The estimated means of Gaussian mixtures neutral faces for K=6
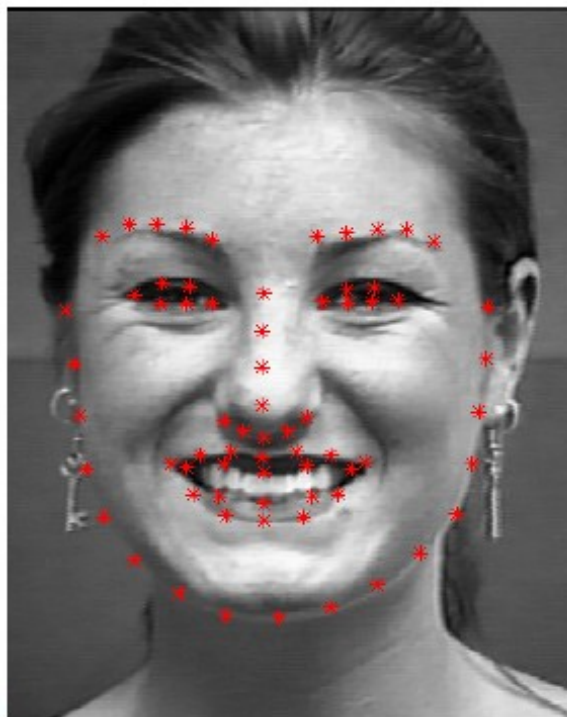
Figure 3.5 (a) : Subject 106 with a happy expression

Source : From CK+ database (© J. Cohn)
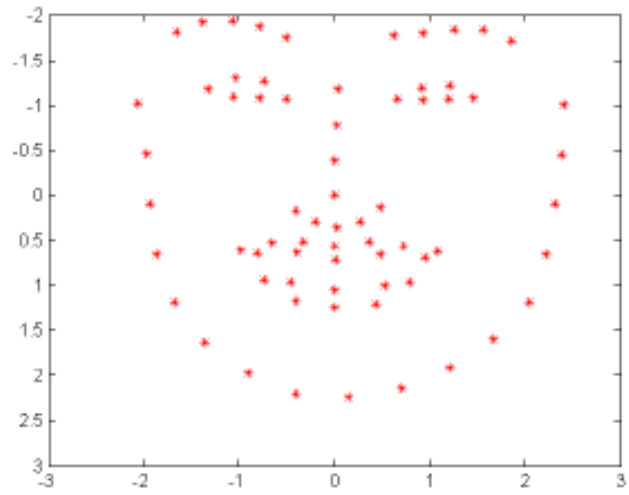
Figure 3.5 (b) : The CBF features after shape alignment
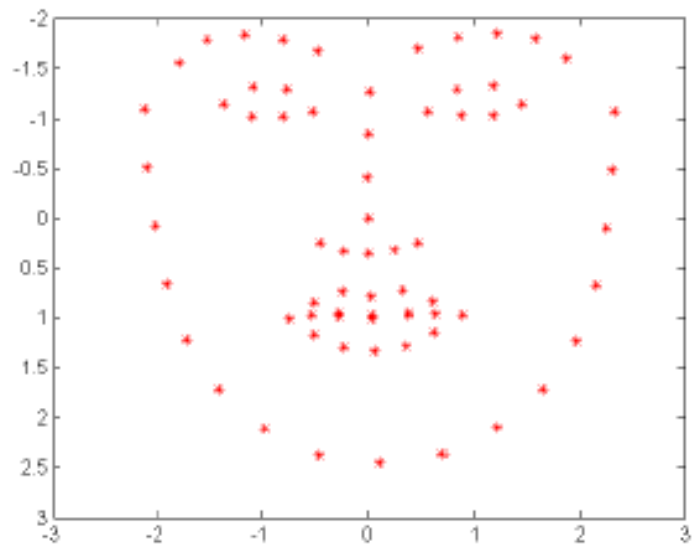
Figure 3.5 (c) : The estimated neutral face shape for the happy expression
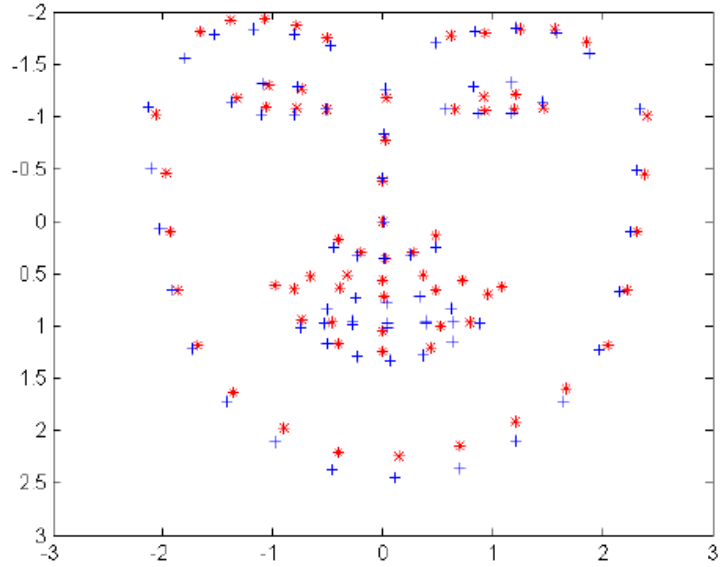
**Figure 3.6 (a) : The CBF features of subject 106
for happy expression (red *) and
its estimated neutral face shape
(blue +) shown together**

We compare the two geometric feature extraction methods (CBF and DABF) under different neutral face shape estimation scenarios. The acronyms corresponding to the compared methods are as follows: **CBF**: Coordinate based features consisting of locations of 68 points, without neutral shape subtraction. **CBF-NS**: Coordinate based features with subtaction of person-specific neutral shape. **CBF-ENS**: Coordinate based features with subtraction of the estimated neutral shape. **DABF**: The 20 distance and angle based features, without neutral shape subtraction. **DABF-NS**: The 20 distance and angle based features, after subtracting the features calculated from the person-specific neutral shape. **DABF-ENS**: The 20 distance and angle based features as Section 3.4, after subtracting the features calculated from the estimated neutral shape.
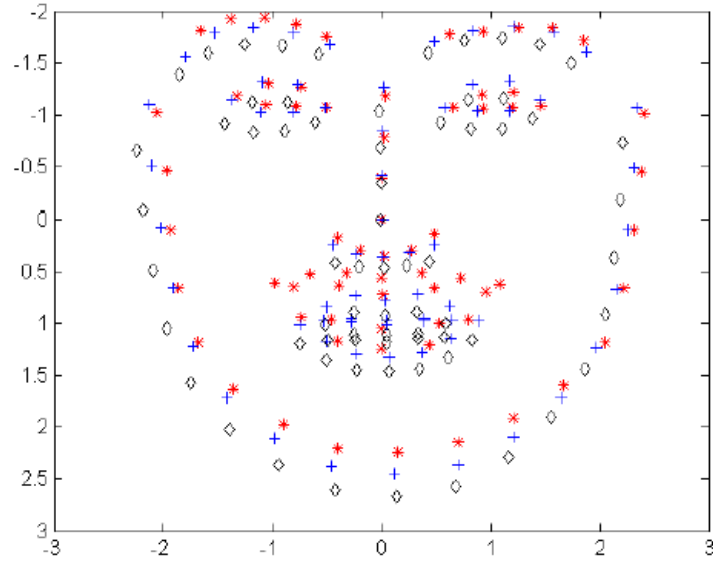
**Figure 3.6 (b) : The happy face shape (red *), the
best fitting neutral face shape
(blue +) and the worst fitting
neutral face shape (gray diamond)**

A Support Vector Classifier with a second order polynomial kernel (Vinciarelli, Pantic and Bourlard 2009) is used to classify the facial features. In order to maximize the training set and to guarantee subject-independence, we use a leave-one-subject-out cross validation scheme. The average emotion recognition rates for the above six facial feature extraction methods are given in Table 3.1. The recognition rate for the proposed CBF-ENS features (88 percent) is higher than the CBF features (83 percent), which shows that estimating the neutral face shape and subtracting it from the shape under test is beneficial. The highest recognition rate is achieved for the CBF-NS (94 percent) features as expected, since person-specific neutral face information is used. We can observe from Table 3.1 that the recognition rates achieved with DABF features are lower than CBF features. However, the proposed neutral face shape estimation method is also beneficial for this feature set, since the recognition rate of DABF-ENS (74 percent) is higher than the recognition rate of DABF features (69 percent). The confusion matrices for the CBF-NS and CBF-ENS features are given in Table 3.2 and Table 3.3, respectively. We can see that happy and surprise have the highest and contempt has the lowest recognition rates. Supporting visual content showing the motion vectors of contempt and surprise can be seen in Figure 3.7.
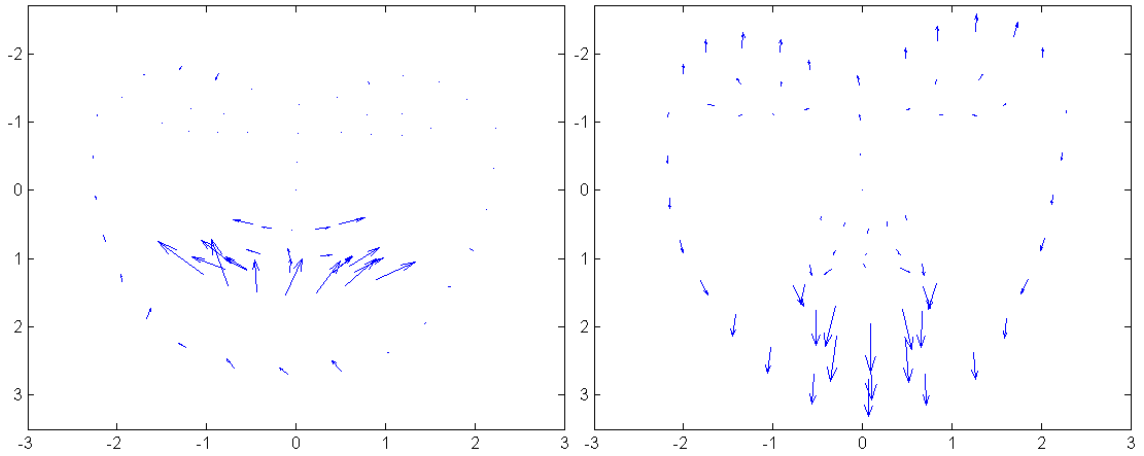
Figure 3.7 : Motion vectors of contempt and surprise emotions

**Table 3.1 The average emotion recognition rates for the six compared feature
sets using a SVC with a second order polynomial kernel**

Features Used	Average Recognition Rate
CBF	83 %
CBF-NS	**94 %**
CBF-ENS	88 %
DABF	69 %
DABF-NS	77 %
DABF-ENS	74 %

Source : Erdem, Ulukaya, and Erdem submitted

Table 3.2 Confusion matrix for CBF-NS features

	An	Di	Fe	Ha	Sa	Su	Co
An	**93.3**	4.44	0	0	0	0	2.22
Di	5.08	**94.4**	0	0	0	0	0
Fe	0	0	**92**	8	0	0	0
Ha	0	0	0	**100**	0	0	0
Sa	3.57	0	0	0	**92.9**	0	3.57
Su	1.2	0	2.41	0	0	**95.2**	1.2
Co	0	0	0	0	5.56	5.56	**88.9**

Source : Erdem, Ulukaya, and Erdem submitted

Table 3.3 Confusion matrix for CBF-ENS features

	An	Di	Fe	Ha	Sa	Su	Co
An	**91.1**	8.89	0	0	0	0	2.22
Di	3.39	**93.2**	0	0	0	1.69	1.69
Fe	0	0	**88**	4	4	4	0
Ha	1.45	2.9	0	**95.7**	0	0	0
Sa	7.14	0	0	0	**89.3**	0	3.57
Su	0	1.2	1.2	0	0	**96.4**	3.57
Co	22.2	0	11.1	0	5.56	0	**61.1**

Source : Erdem, Ulukaya, and Erdem submitted

The average emotion recognition rates achieved with the proposed neutral shape estimation method and coordinate based features is 88 percent, which is higher than the baseline results presented in (Lucey *et al.* 2010), although we do not use the person-specific neutral shapes, and any appearance based features. If we use person-specific neutral face shapes, the recognition rate increases to 94 percent. One comment on the success of recognition rate of our algorithm can be our proposed alignment step. Alignment errors give bad results on the recognition performance since the only

corresponding landmarks are used in the emotion recognition algorithm. In Figure 3.7, one can see that our proposed alignment algorithm is successful to align eye, mouth and nose region which are important in estimating neutral face of the subject. The left image is original points of happy expression and its neutral face shape and the right one is the aligned points of the happy and neutral face shapes. Another reason to get a high performance is to use a polynomial kernel instead of linear kernel in the support vector classifiers as compared to the previous works. Using an SVM classifier with linear kernel as in (Lucey *et al.* 2010) still gives higher recognition rates in our method (88 percent versus 83.3 percent), which shows that the alignment method has a significant effect.

GMM has superior performance than k-means and fuzzy k-means algorithms on estimating neutral face. We first obtained the overall mean of the neutral faces and subtracted it from expressive faces. The average emotion recognition rate is worse (85 percent) than the proposed GMM method (88 percent). We indeed than considered the k-means/fuzzy k-means algorithms. The recognition rates were worse (86.4 percent, 87.3 percent) than the proposed GMM method (88.2 percent), for K=6.
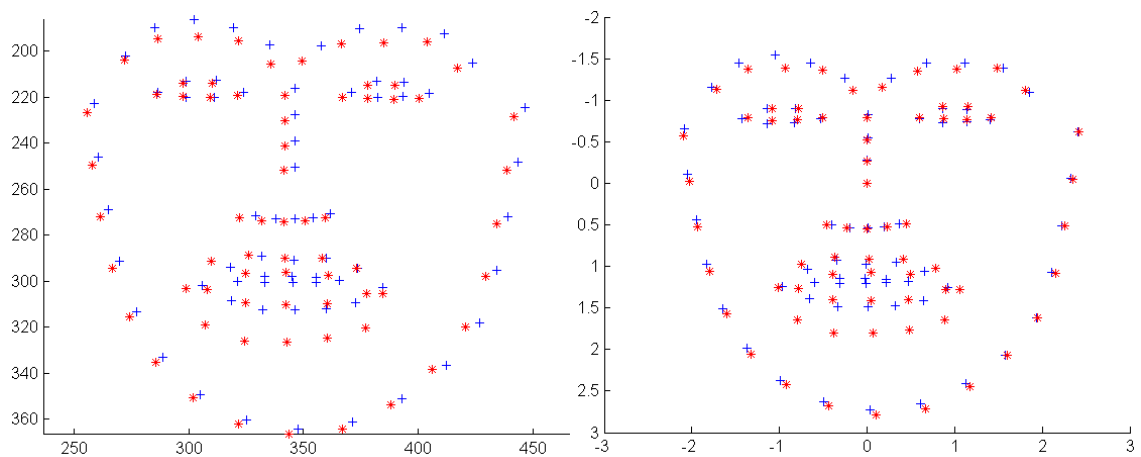

**Figure 3.8 : Effect of the proposed alignment (red * happy expression, blue +
neutral face shape)**

4. FACE RECOGNITION BASED IMDB PLUG-IN APPLICATION

Automatic recognition of cast in movies and to supply information to the viewers is useful when a person watching a movie can not know or recall the celebrities in the movie. In such cases, it is time-saving to find out who the celebrity is using a recognition system instead of searching cast in the Internet. Another application of this plug-in is to search for the identity of a suspicious person quickly, who is observed in a surveillance video.

Face recognition in movies is a difficult task due to variations in illumination, pose, camera position, scale and accessories. In some of the previous works (Li 2007; Ramanan 2007; Arandjelovic 2005; Sivic 2005) face recognition is done only for frontal poses but there are recent works (Fischer 2010; Sivic 2009) that consider other poses as well. The common point of these works is to match faces in facial image sequences. In this work, facial image sequences extracted from the movies is matched with facial image sets gathered from the web. Although face recognition problem has been studied a lot, matching face sequence and face sets case is a challenging problem. Therefore, analysis of the performance of such a matching is important.

Face recognition in movies is a difficult problem but some of the properties of this application make the problem simpler and lead to an increase in the performance. Since the people to be recognized are famous, many training images can be found. Also, by using the available cast information, the number of classes to be matched can be decreased. The developed system starts to function when the user clicks on a face, and stops after presenting the IMDB web page of the celebrity to the user. First, frontal and profile faces are detected using our face detection algorithm around the clicked region. Once the face detected, it is tracked forward and backward in time in subsequent frames. The aim of this tracking is to get other poses of the tracked person. After tracking, local appearance based facial feature vectors are extracted from the facial image sequence. Then, these extracted facial feature vectors are matched with the feature vectors which were extracted offline for the training database. After this comparison, the most

probable three results are returned to the user as the recognition results to direct the user to the IMDB web page of the celebrity.

The system is tested on the Google (2011) image search based database and a face recognition performance of 62 percent is achieved. It must be emphasized that images in the database are not aligned and are not controlled in terms of illumination and occlusion. The remaining part of this section methods and overview of the system will be given in more detail, which are followed by experimental results and conclusion.

4.1 METHODS AND OVERVIEW OF THE SYSTEM

4.1.1 Construction of Training Database

In this part, using Google image search a training database is constructed. Using a script 400 images per celebrity is found. There are 25 celebrities in the database. After that face candidates are detected using the VJ face detection algorithm By visual inspection, non-face images are discarded from the face candidates. Since some of the celebrities have not enough face to construct a database, seven of them are eliminated. After the elimination phase, for only 18 subject, a total of 18x170 face images are obtained. Ten of the subjects are woman.
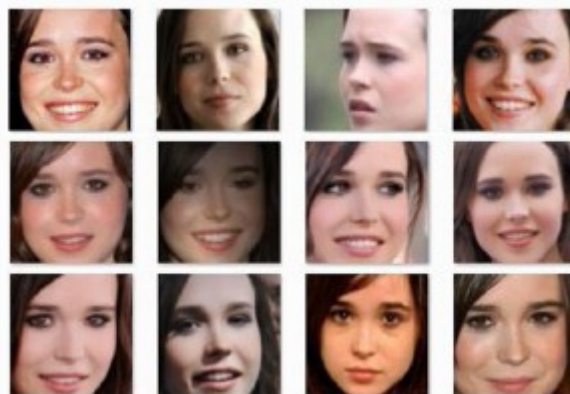

Figure 4.1 : Sample images from the training database
Source : Ellen Page (from Google)

4.1.2 Training the System and Facial Feature Extraction

For the training part some specific facial features must be extracted from the database. For this purpose the 2D Discrete Cosine Transform (DCT) is used as feature extractor.

These found feature vectors are local appearance based feature vectors. The algorithm for the feature extraction is as follows:

1. If it is a color image, convert it to a gray level image.

2. Rescale the image to a 64x64 resolution.

3. Rescaled image is divided into 8x8 blocks. First and last blocks are discarded to reduce background variations and a 64x48 image is obtained.

4. Each block is transformed using 2D DCT.

5. The 8x8 DCT coefficients are ordered in a zigzag manner. A 1x64 vector is obtained.

6. First coefficient (DC coefficient) is omitted from this vector but first five coefficient are retained. For each block a 1x5 local feature vector is constructed. Each local feature vector is divided by its norm for the purpose of normalization to unit length.

7. Finally, each local feature vector of each block are concatenated and a 1x240 (for 48 blocks in one image) global feature vector is composed for each image.

The reason to discard the first coefficient is that it represents the average intensity of the block and does not possess details. The chosen coefficients are distinctive and specific. For example, the second and third coefficients represent vertical and horizontal variations, respectively.

We do unit normalization of local feature vectors to reduce the illumination effects The above procedure is applied to all the images in the database to form the training data. The training data is constructed offline and saved to a file to be used later in the matching task.

4.1.3 Detection of Face Images in Movies

Before face recognition face images are detected and extracted from movies. User clicks on a frame to start the process. After the user clicks on a frame, the VJ face detection

algorithm searches for possible faces in the whole frame. If the VJ algorithm can not find the face where the user clicked, it gives a warning. If it can find the face where the user clicked, face image detection and extraction process starts. The steps of this algorithm are as follows:

1. After the face detector finds the face, different colored rectangles are drawn around the faces regarding their pose.

2. Starting from the clicked frame, the face is tracked forwards and backwards in time. Tracking ends when 50 frames are reached or when the face is lost.

3. The VJ face detection algorithm runs at every frame and detected face regions are saved to a folder.

Face detection algorithm runs until the face is lost. Tracking ends if the face detector can not find a face in three consecutive frames.

The reason to run face detector at each frame is to use the face detector as a face tracker. We assume that the maximum in plane rotation is 15 degrees, and out of plane rotation is 45 degrees. Haar cascades we use are trained to find the profile faces but its performance is low as compared to the frontal one. In order to make the tracking more robust, it is assumed that a face rectangle can not move more than 10 pixels in consecutive frames.

4.1.4 Face Recognition

After the faces are detected, the system is ready for the recognition part. The same feature vector extraction method used in the training phase is also used to extract the features from the face images obtained from the video. In order to compare the test and training data the k-nearest neighbour method is used. The distance metric used is the L1 norm:

$$d = \sum_{z=1}^{Z} \left| f_{training,z} - f_{test,z} \right| \qquad\qquad\text{(4.1)}$$

Each global test feature vector is compared to the vectors in the database. The indices of the most probable three results are returned as the candidates of recognized celebrities. Note that in our system we use a movie-specific cast database instead of a global database to achieve a higher recognition performance.

4.2 EXPERIMENTAL RESULTS

The system is tested on datasets which are constructed using various methods. Different test scenarios are applied to analyze the recognition performance.

In the first scenario, the database images and the test images obtained by Google image search are used in an uncontrolled manner. That is, the detected faces are not verified by visual inspection to see the performance of the VJ algorithm. The results are summarized in Table 4.1.

Table 4.1 Automatic face detection without visual inspection

Number of training images per subject	Number of test images per subject	Recognition rate
150	150	%37.7
200	100	%39.1
250	50	%43.2

Source : Ulukaya, Kayım and Ekenel 2011

In the second scenario, non-face images are discarded from the whole database by visual inspection. The results of this scenario are given in Table 4.2. In this case, the performance increases relative to the first scenario. Ten-fold cross validation is used to get less subjective results. If the number of images per subject in the training set increases, the performance also increases.

Table 4.2 Automatic face detection with visual inspection

Number of training images per subject	Number of test images per subject	Recognition rates with 10 fold cross validation
85	85	%51.6
110	60	%56.2
150	20	**%61.8**

Source : Ulukaya, Kayım and Ekenel 2011

We tested our algorithm on films and movies and observed that the   performance increased when we used movie-specific cast databases. If the tested film was not very popular, the cast images could not be found easily and this decreased the performance and quality of the database. An image of the user interface of the application can be seen in Figure 4.2. The actress Ellen Page is found correctly by this system as seen in Figure 4.2.
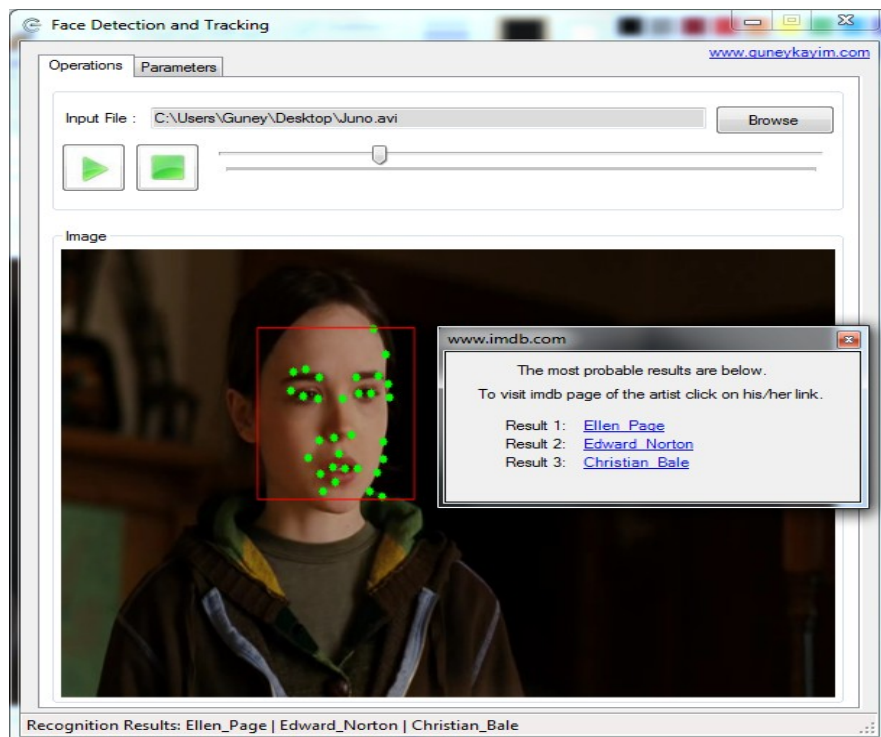

Figure 4.2 : User interface of the plug-in showing the recognition results

Source : Ulukaya, Kayım and Ekenel 2011

4.3 CONCLUSIONS

In this chapter, a face recognition-based IMDB plug-in is presented. When frontal faces are used the recognition results are good but not very high because no alignment procedure was used on the database. One solution to this problem may be to extend the database with various poses of the celebrities. It was observed that using a cast-specific database gives better results.

It is foreseen that this kind of a HCI scenario will become popular in the near future and using an AAM alignment procedure could increase the recognition rate. In movies the statistics of who is appeared in the movie most can be given to user.

Scalability of the proposed algorithm is limited due to the followed algorithm. It is a trade-off between processing time and accuracy. If one uses an alignment step, the accuracy probably will be high but processing time will be high. Since our proposed method has no alignment step, accuracy is low, but processing time is low, too. Also, if the number of subjects is increased, processing time will increase a result of high computational load.

Relevance feedback mechanism could be used to increase the accuracy but it would be infeasible if does not know the celebrity at all. An interesting extension for the automatic attendence taker could be using the relevance feedback mechanism to improve the retrieval performance.

In the early stages of the proposed application, MATLAB (MATLAB 2011),namely MATrix LABoratory, is used to see the primitive results and performance. Then, make the algorithm run faster, OpenCV is used with Microsoft Visual Studio 2011 (Visual Studio 2011). C++ (C++ 2011) software language, namely C Plus Plus, is used to coded the plug-in since it is very fast. Since training is done offline, it is operating near real-time to compare image sets with image sequences.

5. CONCLUSIONS AND FUTURE WORK

In this thesis, first a method for combining the Haar feature based face detector (Viola and Jones 2004) which use brightness information with a skin-color classifier in a post-processing framework is developed. We compared two methods for skin pixel classification: Bayesian method with the histogram technique and the explicit method. We also used an illumination compensation step prior to skin color detection. The experimental results on the Bao (Frischholz 2011) color face image dataset show that the skin-color post-filtering method using the Bayesian classifier is superior to the original VJ (Viola and Jones 2004) algorithm and a pre-filtering method in the literature (Tabatabaie 2009).

Then, a Gaussian Mixture Model (GMM) fitting method for estimating the unknown neutral face shape for frontal facial expression recognition using geometrical features is presented. The distribution of the neutral face shapes in the population is modeled using a GMM with K components, where K is optimized using the Akaike information criterion. Then, the mean vectors of the K Gaussian components represent the typical neutral face shapes in the population. Given a face shape of a facial expression, the "best-fitting" neutral shape is estimated using landmarks that are not affected from facial expressions. The estimated neutral face shape is then subtracted from the face shape under test, to eliminate the person-specific component. The expression related component of the face shape is then classified using a SVC with a second order polynomial kernel.

Experimental results on the CK+ database (Lucey *et al.* 2010), show that estimating the neutral face shape and subtracting it from the landmarks of the test frame is beneficial for increasing the average emotion recognition rate. The average emotion recognition rates achieved with the proposed neutral shape estimation method and coordinate based features is 88 percent, which is higher than the baseline results presented in (Lucey *et al.* 2010), although we do not use the person-specific neutral shapes, and any appearance based features. If we use person-specific neutral face shapes, the recognition rate increases to 94 percent.

We also observed that coordinate based features (Lucey *et al.* 2010) perform better than distance and angle based features (Jiao and Pantic 2010) for the emotion recognition task. It is expected that during the mid-frames recognition performance will decrease. Surprise and happy emotions are the best recognized emotions and sad and angry emotions are the most confused emotions.

One future research direction for improving the face detection in color images by decreasing the false positive rate might be to add an extra validation step based on eye detection. In order to improve the performance of the emotion recognition from facial expressions, appearance based features might be added to the geometric features. It will be interesting to do these experiments on videos with voice, since speech could affect the recognition performance in an adverse manner. Emotion recognition performance on spontaneous databases is also planned to gain insight about posed versus spontaneous cases.

Future research plan for face recognition-based IMDB application could be to develop a new application that automatically takes attendence with a teacher feedback module. Feedback module could be used for validating the attendence results while saving time.

REFERENCES

Books

Alpaydın, E., 2010, "*Introduction to Machine Learning* ", Book, The MIT Press.

Bishop, C. M., 2006, "*Pattern Recognition and Machine Learning*", Book, Springer.

Sonka, M., Hlavac, V., and Boyle, R., 2008, Image Processing, Analysis, and Machine Vision, Third Edition, Thomson Engineering.

Periodicals

Akaike, H., 1974, "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, **19**(6):716 – 723.

Al Haj, M., Bagdanov, A. D., Gonzalez, J., and Roca, X. F., 2009, "Robust and efficient multipose face detection using skin color segmentation", in *Lecture Notes in Computer Science*, vol. **5524**, pp. 152–159.

Arandjelovic, O., Zisserman, A., 2005, "Automatic face recognition for film character retrieval in featurelength films", In: Proc. of the Conference on Computer Vision and Pattern Recognition, pp. 860–867.

Ashraf, A. B., Lucey, S., Cohn, J. F., Chen, T., Ambadar, Z., Prkachin, K. M., and Solomon, P. E., 2009, "The painful face – pain expression recognition using active appearance models". *Image and Vision Computing*, **27**(12):1788–1796

Chen, H. Y., Huang, C. L., and Fu, C. M., 2008, "Hybrid-boost learning for multi-pose face detection and facial expression recognition," *Pattern Recognition*, vol. **41**, no. 3, pp. 1173–1185.

Cootes, T. F., Edwards, G. J., and Taylor, C. J., 2001, "Active appearance models" , *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**(6):681– 685.

Cootes, T.F., Taylor, C.J., Cooper, D., and Graham, J., 1995, "Active Shape Models-Their Training and Application" , Computer Vision and Image Understanding, vol. **61**, no. 1, pp. 38-59.

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

Erdem, C. E., Ulukaya, S., and Erdem, A. T., 2011, "Estimation of the Neutral Face Shape Using Gaussian Mixture Models", submitted to Int. Conf. on Multimodal Interaction, Spain, 2011.

Erdem, C. E., Ulukaya, S., Karaali, A., Erdem, A. T., 2011, "Combining Haar Feature and Skin Color Based Classifiers for Face Detection", *IEEE 36th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, Prague, May 22-27.

Fischer, M., Ekenel, H.K., Stiefelhagen, R., 2010, "Person re-identification in TV series using robust face recognition and user feedback", *Multimedia Tools and Applications*, DOI: 10.1007/s11042-010-0603-2.

Freund, Y. and Schapire, R. E., 1997, "A decision-theoretic generalization of on-line learning and an application to boosting" , *Journal of Computer and System Sciences*, vol. **55**, no. 1, pp. 119–139.

Funt, B., Barnard, K., and Martin, L., 1998, "Is machine colour constancy good enough ?".

Gajsek, R., Struc, V., and Mihelic, F., 2010, "Multi-modal emotion recognition using canonical correlations and acoustic features" , In *International Conf. Pattern Recognition (ICPR)*.

Gunes, H. and Pantic, M., 2010, "Automatic, dimensional and continuous emotion recognition" , *International Journal of Synthetic Emotions*, **1**(1):68–99.

Hjelmas, E., and Low, B. K., 2001, "Face detection: A survey" , *Computer Vision and Image Understanding*, vol. **83**, no. 3, pp. 236–274.

Hsu, R. L., Abdel-Mottaleb, M., and Jain, A. K., 2002, "Face detection in color images", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. **24**, no. 5, pp. 696–706.

Jiao, J. and Pantic, M., 2010, "Implicit image tagging via facial information" , In *ACM Multimedia, Workshop on Social Signal Processing (SSPW'10)*, pages 59–64, Firenze, Italy.

Jones, M. J. and Rehg, J. M., 2002, "Statistical color models with application to skin detection," *International Journal of Computer Vision*, vol. **46**, no. 1, pp. 81–96.

Kakumanu, P., Makrogiannis, S., and Bourbakis, N., 2007, "A survey of skin-color modeling and detection methods," *Pattern Recognition*, vol. **40**, no. 3, pp. 1106–1122.

Kanade, T. J,. Cohn, J., and Tian, Y., 2000, "Comprehensive database for facial expression analysis" , In *Proc. of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, Grenoble, France.

Kim B., Ban, S., and Lee, M., 2008, "Improving adaboost based face detection using face-color preferable selective attention", *Lecture Notes in Computer Science*, vol. **5326**, pp. 88–95.

Li, P., Ai, H., Li, Y., Huang, C., 2007, "Video parsing based on head tracking and face recognition", In: Proc. of the ACM Intl. Conf. on Image and Video retrieval, pp. 57–64.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I., 2010, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression" , In *Proceedings of IEEE workshop on CVPR for Human Communicative Behavior Analysis*, San Francisco, USA.

Niyoyita, J. P., Tang, Z. H., and Ping, L. J., 2009, "Multi-view face detection using six segmented rectangular features", in *The Sixth International Symposium on Neural Networks (ISNN 2009)*, vol. **56**, pp. 333–342.

Ortony, A., and Turner, T. J., 1990, "*What's basic about basic emotions?*", Psychological Review, **97**, 315-331.

Pantic, M., 2009, "Machine analysis of facial behaviour: naturalistic and dynamic behaviour" , *Philosophical Transactions of the Royal Society B-Biological Sciences*, **364**(1535):3505–3513.

Pantic, M. and Rothkrantz, L., 2000, "Automatic analysis of facial expressions: The state of the art" , *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(12):1424–1445.

Phung, S. L., Bouzerdoum, A., and Chai, D., 2005, "Skin segmentation using color pixel classification: Analysis and comparison," *Ieee Transactions on Pattern Analysis and Machine Intelligence*, vol. **27**, no. 1, pp. 148–154.

Ramanan, D., Baker, S., Kakade, S., 2007, "Leveraging archival video for building face datasets", In: Proc. of the Intl. Conf. on Computer Vision, pp 1–8.

Rowley, H.A., Baluja, S., and Kanade, T., 1998, "Rotation Invariant Neural Network-Based Face Detection" , CVPR, pp.38, 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'98).

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

Shobana, L., Yekkala, A. K., and Eajaz, S., 2007, "Face detection using skin segmentation as pre-filter" , in *Int. Conf. on Advances in Pattern Recognition*.

Sivic, J., Everingham, M., Zisserman, A., 2005, "Person spotting: video shot retrieval for face sets", In: Proc. of the Conf. on Image and Video Retrieval, pp. 226–236.

Sivic, J., Everingham, M., Zisserman, A., 2009, "Who are you?"—learning person specific classifiers from video", In: Proc. of the Conf. on Computer Vision and Pattern Recognition, pp. 1145–1152.

Solina, F., Peer, P., Batagelj, B., Juvan, S., and Kovac, J., 2003, "Color-based face detection in the 15 seconds of fame art installation" .

Tabatabaie, Z. S., Rahmat, R. W., Udzir, N. I. B., and Kheirkhah, E., 2009, "A hybrid face detection system using combination of appearance-based and feature-based methods" , *Int. J. of Computer Science and Network Security*, vol. **9**, no. 5.

Ulukaya, S., and Erdem, C. E., 2011, "Duygu Tanıma için Geometrik Yüz Özniteliklerinin Karşılaştırılması", *IEEE 19th Signal Processing and Applications Conference* (SIU), Antalya, Turkey, April.

Ulukaya, S., Kayim, G., and Ekenel H. K., 2011, "Filmler için Yüz Tanıma Tabanlı IMDB Eklentisi", *IEEE 19th Signal Processing and Applications Conference* (SIU), Antalya, Turkey, April.

Vezhnevets, V., Sazonov, V., and Andreeva A., 2003, "A survey on pixel-based skin color detection techniques," *Proc. Graphicon*.

Vinciarelli, A., Pantic, M., and Bourlard, H., 2009, "Social signal processing: Survey of an emerging domain". *Image and Vision Computing*, **27**(12):1743-1759.

Viola, P. and Jones, M. J., 2004, "Robust real-time face detection," *Int.J. of Computer Vision*, vol. **57**, no. 2, pp. 137–154.

Viola, P. and Jones, M., 2001, "Rapid object detection using a boosted cascade of simple features".

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

Wang, X., Xu, H., Wang, H., and Li, H., 2008, "Robust real-time face detection with skin color detection and the modified census transform" , in *Proc. of IEEE Int. Conf. on Information and Automation*, China.

Wu, B., Ai, H. Z., Huang, C., and Lao, S. H., 2004, "*Fast rotation invariant multi-view face detection based on real Adaboost*", Sixth IEEE Int. Conf. on Automatic Face and Gesture Recognition.

Yang, M. H., Kriegman, D. J., and Ahuja, N., 2002, "Detecting faces in images: A survey", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. **24**, no. 1, pp. 34–58.

Zeng, Z. H., Pantic, M., Roisman, G. I., and Huang, T. S., 2009, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions" , *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**(1):39–58.

Zhang, C., and Zhang, Z., 2010, "A survey of recent advances in face detection", Technical Report, Microsoft Research.

Other Publications

C++. 2011, Available from: http://www.cplusplus.com/doc/tutorial/ (cited May 01,2011).

eNTERFACE'05. 2011, "Emotional database", Available from: http://psy.ff.uni-lj.si/Katedre/PM/speech_emotion/clanki/Audiovisual%20emotion%20database.pdf (cited May 01, 2011).

Fasel, I., Dahl, R., Hershey, J., Fortenberry, B., Susskind, J., and Movellan, J., 2004, "The machine perception toolbox", http://mplab.ucsd.edu/grants/project1/free-software/mptwebsite/API/, (cited May 01, 2011) .

Google. 2011, Available from: http://www.google.com/imghp (cited June 01, 2011).

Humaine. 2011, Available from: http://emotion-research.net/ (cited May 01, 2011).

iCub. 2011, Available from: http://en.wikipedia.org/wiki/ICub (cited May 01, 2011).

IEMOCAP. 2011, Available from: http://sail.usc.edu/iemocap/ (cited May 01, 2011).

JAFFE. 2011, Available from: http://www.kasrl.org/jaffe.html (cited May 01, 2011).

MATLAB. 2011, Available from: http://www.mathworks.com/products/matlab/ (cited May 01, 2011).

MMI. 2011, Available from: http://www.mmifacedb.com/ (cited May 01, 2011).

OpenCV. 2011, Available from: http://sourceforge.net/projects/opencvlibrary/ (cited Jan 01, 2011).

PRTools. 2011, The MATLAB toolbox for pattern recognition, Available from: http://www.prtools.org/index.html, (cited May 01, 2011).

Robotics. 2011, Available from: http://www.mbhs.edu/~lpiper/Robotics03/develop.html (cited May 01, 2011).

R. Frischholz, 2011, "Bao face database at the face detection homepage", in *http://www.facedetection.com* , (cited Oct 05, 2010).

Ulukaya, S., 2011, "Ground truth data for Bao face database", Available from: http://staff.eng.bahcesehir.edu.tr/~cigdemeroglu/groundTruthBAO.rar (cited May 01, 2011)

Visual Studio. 2011, Available from: http://www.microsoft.com/download/en/ (cited May 01, 2011).

CURRICULUM VITAE

FULL NAME	**: Sezer Ulukaya**
ADDRESS	**: Bahçeşehir University Faculty of Engineering**
BIRTH PLACE / YEAR	**: Malatya / 1985**
LANGUAGE	**: Turkish (native), English**
HIGH SCHOOL	**: Malatya Anatolian High School**
B. Sc	**: Electronics Engineering, Ankara University**
M. Sc	**: Electrical and Electronics Engineering, Bahçeşehir University**
NAME OF INSTITUTE	**: Natural and Applied Sciences**
NAME OF PROGRAM	**: Electrical and Electronics Engineering, (Embedded Video Systems - Video Track)**
WORK EXPERIENCE	**: From September 2008 – Currently Bahçeşehir University, Teaching Assistant**