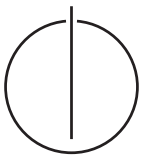
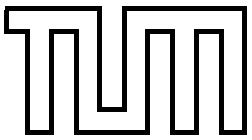


Institut für Informatik
Technische Universität München



Visual Interpretation of Human Body Language for Interactive Scenarios

Dissertation

Zahid Riaz

Lehrstuhl für Bildverarbeitung und Mustererkennung
Institut für Informatik
Technische Universität München

Visual Interpretation of Human Body Language for Interactive Scenarios

Zahid Riaz

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Hans Michael Gerndt
Prüfer der Dissertation:	1. Univ.-Prof. Michael Beetz, PhD
	2. Prof. Dr. Mubarak Shah, University of Central Florida, USA

Die Dissertation wurde am 09.06.2011 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 08.11.2012 angenommen.

Abstract

Since the recent advancements in the area of interactive technologies real time analysis and interpretation of the body language has become a challenging problem for many researchers. Furthermore, interpretation of human actions and facial behaviors for perception learning plays a significant role in *Human Robot Interaction* (HRI) scenarios. A markerless and automatic analysis of the body language helps the robots to jointly perform activities with humans by extracting the distinctive set of features. This thesis investigates two aspects of the visual body language analysis for real world scenarios. These two aspects are face image analysis and full body action recognition for interactive systems.

For human face image analysis, 2D and 3D modeling techniques are studied in detail showing that better results are obtained by using a 3D face model. The proposed approach utilizes context information and is robust against varying poses and facial expressions for face recognition applications. One of the major contributions of this thesis for face image analysis is the extraction of a *Spatiotemporal Multiple Feature* (STMF) from the image sequences. These features show their strength in unconstrained analysis of the face images and have been successfully used for face recognition, facial expressions classification, gender classification and age estimation. An STMF consists of facial structural and textural parameters along with temporal parameters extracted from the local feature motions. A special attention is paid to facial appearance which includes image registration, texture representation and finding texture descriptors. The proposed feature set outperforms current state of the art feature extraction approaches for facial classification which reflects their high performance, cost effectiveness and integration flexibility to real world systems.

For full body actions recognition, *bag of words* (BoW) approach is studied which utilizes image contents. This approach is studied for full body actions and facial expressions recognition. This is a markerless approach which does not require any prior knowledge about object detection, person tracking and initial position. Different variants of this approach have been studied in this thesis. A thorough analysis for optimal choice of spatiotemporal features has been conducted in comparison to original the BoW approach. The experimental results show that this approach outperforms context based modeling of the full body for activity recogni-

tion. In order to validate the representation strength of the proposed feature vectors, extensive experiments have been conducted on various standard databases. These databases are widely used for benchmarking purposes and provide real world variations for developing new algorithms.

Kurzfassung

Durch die jüngsten Entwicklungen im Bereich der interaktiven Technologien wurden die Analyse und die Interpretation von Körpersprache in Echtzeit eine anspruchsvolle Problemstellung für viele Forscher. Des Weiteren spielt die Interpretation menschlicher Aktivitäten und Gesichtsausdrücke eine wichtige Rolle im Kontext der Mensch-Roboter-Interaktion (MRI). Eine markerlose, automatisierte Analyse von Aktionen und Gesichtsausdrücken hilft Robotern dabei, zusammen mit Menschen gemeinsame Aktivitäten durchzuführen, indem sie eine nützliche Menge an Merkmalen extrahiert. Diese Arbeit beschäftigt sich mit zwei wesentlichen Aspekten der Körpersprache: der Analyse von Gesichtern und der Erkennung von Ganzkörperaktionen für interaktive Systeme.

Für die Analyse des menschlichen Gesichtes wurden Modellierungsansätze in 2D und 3D im Detail verglichen, wobei durch Verwendung der 3D Gesichtsmodelle bessere Ergebnisse erzielt wurden. Der hier vorgeschlagende Ansatz verwendet Kontextinformationen und ist robust gegenbüber Änderungen in der Pose sowie Änderungen im Gesichtsausdruck. Einer der Hauptbeiträge der Arbeit liegt in der Extraktion von verschiedenen räumlichen und zeitlichen Merkmalen aus Bildsequenzen und in der Beschreibung dieser Merkmale in einer Merkmalsmenge (STMF). Diese repräsentativen Merkmale zeigen ihre Stärke besonders bei der uneingeschränkten Bildanalyse von Gesichtern. Sie wurden für die Gesichtserkennung, die Klassifikation von Gesichtsausdrücken, die Klassifikation des Geschlechts und die Schätzung des Alters erfolgreich eingesetzt. Eine STMF Merkmalsmenge besteht aus Parametern, die die Struktur und die Textur beschreiben, sowie aus Parametern die Bewegung lokaler Merkmale über die Zeit beschreiben. Ein besonderes Augenmerk wurde auf die Erscheinungsform von Gesichtern gelegt. Dies spiegelt sich in der Bildregistrierung, der textuellen Repräsentation und auch deren Beschreibung wider. Mit der vorgeschlagenen Merkmalsmenge übertrifft andere Ansätze im Hinblick auf Performanz, Kosteneffizienz und Flexibilität zur Integration in realistischen Systemen.

Im Hinblick auf Ganzkörperaktionen wurde der *bag of words* (BoW) -Ansatz basierend auf Bilddaten betrachtet. Dieser Ansatz wird für Ganzkörperaktionen und Gesichtsausdrücke untersucht. Dieser markerlose Ansatz benötigt keinerlei Vorwissen in Bezug auf Personen-

erkennung, Personenverfolgung oder die Initialpose. In der Arbeit wurden verschiedene Varianten des Ansatzes experimentell untersucht. Zur Feststellung einer optimalen Auswahl an räumlichen und zeitlichen Merkmalen wurden diese Varianten mit der ursprünglichen Version des "Bag of Words"-Ansatzes verglichen. Die experimentellen Ergebnisse zeigen dass dieser Ansatz genauer ist als Kontext-basierte Ganzkörpermodellierung für Aktionserkennung. Um die Ergebnisse der Arbeit zu verifizieren, wurden detaillierte und umfangreiche Experimente auf der Basis von standardisierten Datenbaken durchgeführt, welche häufig zu Benchmarkingzwecken verwendet werden.

Acknowledgments

I feel honor to write a few words about some of the precious people and the organizations around me because without their presence it was not possible to achieve this milestone of my life.

Firstly, I would like to thank Prof. Michael Beetz who always motivated me throughout my PhD research and always supported my ideas and timely added his valuable opinions. I really appreciate the *Intelligent Autonomous Systems* (IAS) group established by him and wish this group a very best of luck. Prof. Bernd Radig, one of the most important names in my career who gave me the chance to join TU Munich and his kindness, support and motivations have been always been very appreciative.

I would also like to extend my special thanks to Prof. Mubarak Shah, head of computer vision lab at University of Central Florida (UCF). It was a great experience to work under his supervision and a short exchange experience in his lab has opened several new horizons of computer vision to me. Furthermore, I would like to thank Prof. Michael Gerndt for his support to serve as a chair of the examination committee and his prompt help at the right time to organize everything.

The cooperation of my colleagues working at IAS group and UCF computer vision lab has also been very appreciative, I would like to thank them all including Francisco, Murat, Michael, Karinne, Dr Alexandra, Johannes, Alexis, Federico, Dejan, Dr. Humera, Sikander, Martin, Sabine, Manuela and Quirin. Further, I would say thanks to Lars Kunze and Dr. Haris for their valuable suggestion in writing this thesis and Kishore (UCF) for scholarly discussions at UCF. A special thanks to Christoph Mayer, Dr. Jan Bandouch, Dr. Matthias Wimmer and particularly to Dr. Suat Gedikli for supporting my ideas and helping me to improve my ideas by adding their valuable suggestions. Moreover, the support by Mr. Rainer Worst (Fraunhofer Gesellschaft) during the last phase of my PhD is highly appreciable. I would also like to thank Shahid Riaz, my brother for helping and supporting me to finish this task. Also thanks to three little angels Maaz, Musfera and Rahima who always have many wonderful ideas and will surely be shinning stars of future.

Higher Education Commission (HEC) in Pakistan deserves a big applause here for giving

me a chance and providing me the sufficient funds to pursue my PhD degree. Furthermore, *Cognition for Technical Systems (CoTeSys)* has provided an excellent platform to implement new ideas. In the later stages of my PhD, TU Munich Graduate School (TUM-GS) has flourished my skills and provided me with a chance to collaborate with international research groups.

Munich, October 2012

To my mother
Late Shamim Riaz

Contents

List of Figures

List of Tables

List of Symbols

Symbols related to a general face model.

p	Model parameter vector
$\mathbf{p}^*$	Ground truth model parameter vector
$ii(x, x)$	Integral image at point (x, y)
x, y, and z	x-, y-, and z-coordinate
s	Scaling factor
r	Rotation matrix
t	Translation vector
ϕ	3D rotation matrix
ϕ_α, ϕ_β and ϕ_γ	Rotation matrices around horizontal, vertical and camera axis respectively
x_1, x_2, and x_3	Vertices of a triangle in the face mesh
$\mathbf{b}_s$	Structural parameters of the model
$\mathbf{b}_g$	Textural parameters of the model
$\mathbf{b}_t$	Temporal parameters of the model
u	An *Spatiotemporal Multiple Feature* (STMF)
m	Number of structural parameters
n	Number of textural parameters
p	Number of temporal parameters
(u, v)	Texture coordinates in a texture map

General symbols related to feature extraction and parameterization.

$\overline{X}$	Mean vector
C	Covariance matrix
P	Matrix of eigenvectors
$\phi_1, \ldots, \phi_n$	n eigenvectors of covariance matrix C
b	Parameter vector from PCA
c(u,v)	Correlation coefficient at pixel (u, v)

$\gamma(u, v)$	Normalized correlation coefficients at pixel (u, v)
k	Number of clusters in k-means clustering
$f(I, p)$	An objective function for image I with model parameters p
$W(x; p)$	Piecewise affine warp with position x and pixel value p
H	Homography matrix

Symbols related to action and activity recognition.

σ	Variance of Guassian filter
τ	Temporal parameter of Gabor filter
h_{ev} and h_{odd}	Even and odd components of a Gabor filter
σ_x and σ_y	Variance of Guassian filter in x and y direction

List of Abbreviations

AAM	Active Appearance Model
ASM	Active Shape Model
AU	Action Unit
AUV	Action Unit Vector
BDT	Binary Decision Tree
BN	Bayesian Network
BoW	Bag of Words
CKFED	Cohn Kanade Facial Expressions (CKFE) database
DCT	Discrete Cosine Transform
DT	Decision Tree
FACS	Facial Action Coding System
FPR	False Positive Rate
GUI	Graphical User Interface
HRI	Human Robot Interaction
ICIA	Inverse Compositional Image Alignment
LBP	Local Binary Patterns
MAE	Mean Absolute Error
MMI	Man Machine Interaction
PCA	Principal Component Analysis
PDM	Point Distribution Model
PIE	Pose Illumination and Expressions
ROC	Receiver Operating Characteristics
STMF	Spatiotemporal Multiple Feature
SVM	Support Vector Machine
TPR	True Positive Rate

CHAPTER 1

Introduction

Since the recent advancements in the area of interactive technologies, automatic and markerless analysis of human body language in unconstrained environments has become a challenging task for the researchers. Several systems installed in our daily lives are capable of predicting human activities in real time. Examples include *Microsoft kinect sensor* for human actions recognition [Mic], smart cameras with PC inside which contain operating systems and processing capabilities [Xim], daily life interactive systems like vending machines commanded by human faces [Co.] and *assistive robotics* [Sys], where humans are assisted by the robots to perform their routine tasks. These systems are developed on the baseline principles of artificial intelligence and have capability to quickly adapt to their surroundings. The non-intrusiveness, high performance, efficiency in analysis and generalization properties make these systems capable for installing in daily used systems. These capabilities are realized in the intelligent systems by using tools and techniques from inter-disciplinary cognitive and computational science. Due to their generalization property, such systems have the ability to interact with previously unseen persons without any prior information about the interacting person. These characteristics are obtained for the intelligent systems by making them robust against current outstanding challenges in the visual analysis of body language. Some of these outstanding design challenges for computer vision scientists include varying poses, dynamic facial expressions, changing illuminations, partial occlusions and inter-personal action variations. The goal of this thesis is to develop an automatic body language analysis system which mainly comprises of face image analysis and full body action recognition being robust against aforementioned challenges.

This thesis provides the implementation of two different aspects of visual body language analysis system, which include interpretation of facial behaviors and human actions for *interactive scenarios*. In interactive scenarios, humans and intelligent systems perform together to accomplish a task. We use the terminology of *intelligent systems*, *intelligent machines*, *assis-*

tive systems or *robots* with the similar meaning. Moreover, with *assistive scenarios* we mean the scenarios where robots are helping the humans to perform different activities. This research work is conducted by implementing different new ideas by using a single camera. The overall research approach aims toward the development of a stand-alone system which performs different tasks from preprocessing of the videos/image sequences, feature extraction and object representation, analysis, interpretation to finally feature classification. For face image analysis, we use conventional image based approaches [ZCPR03] with image registration, 2D model based approaches [CET98b] and 3D face modeling from a single image [RGBR10] and compare the results with current state of the art approaches. Our goal is to extract a *Spatiotemporal Multiple Feature* (STMF) which is robust and sufficient to describe different facial attributes. The proposed feature vector is extracted from model based approaches and has been tested in the presence of varying head poses and facial expressions. This feature set is found robust against these variations. In case of full body action recognition, we use sparse spatiotemporal features which are extracted using the famous *bag of words* (BoW) approach. BoW is used with additional spatial features extracted from local descriptors. These features are capable to perform in the presence of varying poses, occlusions and require no context information.

In an abridged form, interpretation of body language describes biometric and soft biometric traits of the humans which include facial identity, facial expressions, gender, estimation of the age and daily life actions and activities performed by the humans. Such systems are ideal for use in *Human Robot Interaction* (HRI) scenarios. Assistive robots, which are designed to help elderlies to perform their daily life activities safely and with privacy can adapt themselves quickly to a new person by acquiring different aforementioned attributes of human body. Since these systems are person independent, they can be further useful for generalized applications like medical care of the patients. The non-intrusive nature and cost effectiveness of these systems recommend to embed them in daily used camera mounted systems to facilitate the humans in assistive environments. The main goal of this thesis is to propose a feature vector extracted from camera captured images which is capable to perform under real world variations.

1.1 Aims and Motivation

The aim of the research work conducted in this thesis is the development of an autonomous system for analysis and interpretation of body language during interaction with the robots. As cameras are getting widespread use and are mounted on computer screens, embedded into daily use mobile phones and installed into everyday living and working environment they

become valuable tools for human system interaction. An important aspect of this interaction is detection and recognition of the faces and interpretation of the facial expressions. These capabilities are deeply rooted in the human visual system and a crucial building block for social interaction. Consequently, these capabilities are an important step toward the acceptance of many technical systems. Most of the systems in our daily lives are mounted with cameras and require automatic analysis of the human actions to predict human intentions. This leads to the autonomous systems capable of interacting intelligently with the humans in unconstrained environments by using prior knowledge and context information to process the information heuristically, understand human intentions and manipulate the tasks in joint activities between humans and the robots. In contrast, to industrial robots, such systems find their applications as medical care robots and assistive robots.

Most of our daily life interactions are face to face with others which help us to predict facial expressions and behaviors of the interacting person easily. Facial emotions provide an initial knowledge about the interaction to manipulate human actions. For example, a handshake between two persons starts from face to face interaction and a pleasant facial expression may lead to a handshake. These goals are accomplished by performing human face recognition along with other soft-biometric traits like gender classification, facial expressions and age estimation. Besides different facial traits, full body actions play a significant role under such scenarios. Hand gestures are often used as command to a system while one or more continuous full body actions form an activity. These actions can be learned by the robots and synthesized to perform similar actions later in assistive scenarios.

1.2 Problem Description

Interpretation of the body language in assistive environments using a single camera is a challenging problem for most of the computer vision researchers. Besides several common challenges from pattern classification perspectives, there are a few issues which are specific to the given problem. Therefore, we separately discuss face image analysis and full body action recognition for a visual body language analysis system.

1.2.1 Face Image Analysis

Human faces are mostly observed in actions conveying a sets of meaningful information which originate most of the daily interactions. Despite the research of past few decades, visual analysis of the faces remains a challenging problem. Although many commercially available sys-

tems implement different facial applications efficiently, the problem still needs to be addressed in the presence of real world challenges like facial poses, illuminations and lightings, varying facial expressions, facial aging, low resolution images and sensor noise, occlusions and especially spoofing. Figure 1.1 describes some of these challenges in face recognition systems. We study different aspects of human faces which are the building blocks of a common human-human interaction. These aspects include biometric and soft biometric attributes of human faces including person identity, facial expression and behavior, gender, age estimation and somehow ethnic origin. Though conventional image based approaches are efficient and exhibit high accuracy under controlled conditions but produce unsatisfactory results when subjected to real world variations. The goal of this research work is to find such a feature set which is robust against current outstanding challenges in the analysis of face images and at the same time can represent different facial characteristics. These facial characteristics are face recognition, facial expressions classification, gender classification, ethnic origin and estimation of the age. This feature set is extracted by using model based approaches. In such applications an automatic and efficient feature extraction technique is required to classify different biometric attributes from the given face images. Furthermore, the extracted feature set should be robust enough to directly apply in real world applications. Most of the currently available systems lack this property. A major reason is that most of the research work recommends to isolate the sources of variations while focusing on a particular application. For example, in face recognition applications, many researchers normalize the faces in order to remove facial expressions variations to stabilize face recognition results against unwanted facial deformations [LJ05]. In these cases, the extracted features do not contain facial expressions information and cannot be used to further classify soft biometric attributes of the faces. We develop a unified feature set which can be used for different applications in the presence of different variations. This is facilitated by using 2D and 3D face models. From the experimental results, it can be seen that the model based approaches describe such systems in a compact manner, while 3D face modeling outperforms 2D models because of better texture realization. We pay a special attention to texture realization and textural feature extraction.

The goal of the model based facial feature extraction is to study an interactive scenario for socially inspired robots. In this scenario, we may come across different people in social gatherings, meetings, conferences and markets. While glancing at a single face, human are able to predict age, gender, ethnicity, facial expressions and identity if the person is already known before. We term this scenario as *face-at-a-glance* and devise a solution for human robot interaction domain. The proposed feature set enables the robots to perform a human-like behavior in social activities. For instance, an autonomous robot serving a group of different persons can

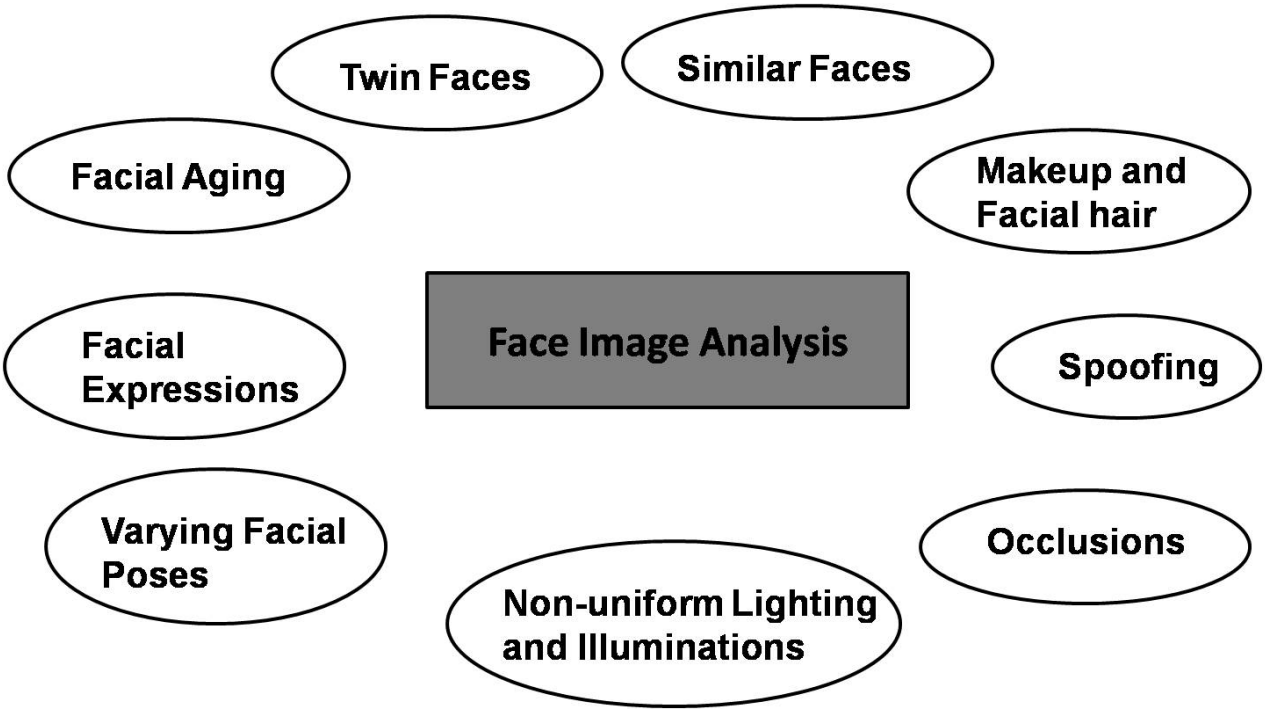


FIGURE 1.1 Some of the challenges in face image analysis like aging effect, twin faces, similar faces, makeups effects, spoofing, occlusions, varying lighting, varying poses, facial expressions.

better interact by finding their face information at a first look. Once a robot knows the gender and age, it can decide from a knowledge base that how to interact with this person. A pleasant human facial expressions may motivate the robot to interact further and by knowing the identity of this person, the robot can perform person specific services. Nevertheless, additional ethnic information can further teach the robot that how to perform inter-cultural interactions.

In addition to challenges in computer vision areas, by their nature human faces exhibit a few challenges which make them harder to classify. These challenges include:

- **Uniqueness:** Human faces are not unique by nature as compared to other biometrics like fingerprints and iris pattern. For example classification between blood-relations and especially twin faces is hard.
- **Performance:** It is harder to classify faces as compared to fingerprints and iris images. Their performance is normally below automated fingerprints identification system (AFIS) in biometric industry.
- **Circumvention:** Faces can be spoofed or a high false positive rate may lead to system failure.

FIGURE 1.2 Face-at-a-glance scenario: Humans are quite efficient in extracting features holistically to interpret face images. With a glance on any face, humans can tell the gender, estimation of the age, predict the ethnic origin, classify facial expression and identity if already known before. We solve the same problem for robots to extract these information using a robust spatiotemporal feature set.

- **Temporal Deformation:** Aging effects are far prominent in faces as compared to fingerprints and iris images.

- **Human visual limitation:** Experiments from cognitive sciences have proved that computer vision system performs better than human vision system for previously unknown faces [O'T09].

1.2.2 Full Body Actions and Activities Recognition

In addition to facial behavior analysis, actions and activities recognition help the robots to understand human motion in order to manipulate the tasks in joint interactions. Automatic recognition of human activities has been addressed by the researchers in several different ways [WRB10]. In HRI scenarios activities are performed in a sequence of seamless actions. This makes the problem harder as one has to decide that when does a task start and at which point does it finish? This causes the task segmentation problem challenging. Another factor is the inter-person variability in performance of the tasks. For instance, same task may be performed by different persons by using one hand (left or right) or two hands. Nevertheless, different body sizes may also decrease the classification rate. Parallel tasking may also lead to misclassification of the task. Figure 1.3 shows these challenges in detail.

FIGURE 1.3 Challenges in action recognition [TBB09]: In assitive environments activities are performed differently by different persons, varying body sizes, knowledge when an action starts and when it finishes and finally, occlusions. We show that bag-of-words approach is robust against some of these challenges.

Moreover, classification of actions and activities is harder in the presence of multiple persons and objects in a scene, cluttered backgrounds, partial occlusions and self occlusions, varying poses and illuminations. Since actions are defined as the movement performed in a sequence in a given context, therefore temporal component also plays a significant role in classification and utilized to overcome some of these challenges. For most of the motion features moving backgrounds cause deterioration in the classification results. We use *bag of words* (BoW) approach to classify different actions. This approach is capable to deal with partial occlusions, varying poses and require no prior knowledge about object detection, tracking or segmentation. It utilizes content information from the video data.

1.3 Contributions

In order to develop an automatic system for body language analysis, we study two individual systems; face image analysis and full body action recognition in unconstrained environments. We focus on assistive scenarios where the humans and the robots perform different tasks together. To achieve these goals, following contributions are made in this thesis:

- ***Spatiotemporal Multiple Feature (STMF):*** In HRI scenarios faces are generally observed in actions performing different dynamics under varying head poses. Human are quite efficient in extracting multiple information from the faces of the interacting persons in a short time. This set of information contains facial expressions recognition, gender classification, age estimation, prediction of the ethnic origin and identity of the interacting person very quickly. We obtain a similar performance for the robots in real world scenarios. This goal is achieved by using human face modeling. After a thorough analysis and experimentation from a 2D point distribution model and a 3D wireframe model of the faces, we come to the conclusions that a single feature set can represent these attributes while being capable to apply in daily life systems. The details about feature extraction from two different models is discussed in Chapter 4 and Chapter 5. This feature set consists of structural, textural and temporal information of the fiducial points which describe a face model. We term this feature vector as *Spatiotemporal Multiple Feature* (STMF) .

  This feature set is one of the major contributions of this thesis for face image analysis. An STMF has been extensively experimented and found stable against varying poses and facial expressions for a face recognition system and at the same time successfully classifies human faces, facial expressions, gender and estimates age. The results are obtained on different databases to validate the representation strength and richness of this feature set. Table 1.1 shows the significance of three different types of parameters with their primary and secondary contributions to this feature set. These results are verified later in Chapter 4 and Chapter 5.

- ***Effects of Perspective Distortions:*** We study face recognition problem under varying view-points and facial expressions by using 3D face modeling techniques. While faces are captured in actions, some of the facial areas which are away from camera are under self occlusions and considered as missing information. However, under limited out-of-plane rotations most of the facial areas are tilted and have very less textural information due to their oblique shape. In conventional face modeling approaches [CET98b], texture

	Identity	Expressions	Gender	Age	Ethnicity
Structural	*p*	*p*	*p*	*p*	*p*
Textural	*p*	*s*	*p*	*p*	*p*
Temporal	*s*	*p*	*s*	*na*	*na*

TABLE 1.1 Contribution of different components to our proposed STMF set, p = primary contribution, s = secondary contribution, na = not applicable. This configuration is provided by a thorough literature survey. The practical evidence of this representation is experimented in Chapter 4 and Chapter 5.

.

distortions at these areas of the face edges are ignored by assuming that the distance between the camera and the face is larger than the original face size. In these cases, effects of perspective distortions are ignored in face recognition application. However, we study this problem in detail and obtain an improved performance by considering perspective effects on these areas. We experimented this approach and obtain better performance on the PIE database [TBB02] which contains pose, illumination and expression variations. This observation provides a significant improvement in face recognition results. It can be concluded that the effect of perspective distortions should not be ignored in case of texture warping.

- ***Expressions and View Invariant Face Recognition:*** Since face image analysis problem is studied on both 2D and 3D face models, we conduct extensive experiments and study the robustness of the face recognition system against facial expressions and varying poses. The extracted feature set (STMF) is a holistic representation of the local variation of the faces. In case of 2D appearance model, we verify that our feature set can represent face recognition, facial expressions recognition, gender classification and age estimation. However, robustness against varying view points is limited. This issue is further improved using a 3D wireframe model. We study that a rendered 3D face model has the ability to deal with pose, illumination and expressions (PIE) invariance in real-world applications. We find that extracting structural, textural and temporal variations from a given image sequences or a video is sufficient to extract multiple information about the face images under different variations. Further, we study compactness of the textural feature vector by using different descriptors. We comparatively use *Local Binary Pattern* (LBP) , *Discrete Cosine Transform* (DCT) and *Principal Components Analysis* (PCA) features and find that PCA performs better with 3D face model because texture is

rendered in a block wise undistorted texture map. In order to assure the compactness of the texture features, we also define an energy based feature which is extracted from each block of the texture map. This energy based feature vector outperforms PCA because of the rectified texture map representation and do not require subspace learning for feature extraction. A significant improvement using energy based feature as compared to PCA has also been reported in this thesis.

- ***Sparse Representation for Face Recognition:*** Besides human face modeling, we develop a face recognition by using conventional image based approaches. An input image is normalized during preprocessing stage and various features are extracted for classification. A basic sequential flow of this process is shown in Figure 1.4. This stand-alone system is capable to work under limited lighting variations, in-plane and out-of-plane rotations and can deal with face localization problems. This preprocessing phase refines the input image and registers all the images to a standard template. Further, facial features localization is also used for facial expressions recognition. We apply several distance constraints on facial features like distance between eyes, relative positions of eyebrows between two frames to classify facial expression in real time. Similar image registration approach has also been used by other researchers [Eke09]. We study different feature extraction approaches on these normalized images. These approaches include holistic and sparse spatial features for face recognition. We also propose an interest point detector by using content based image description techniques. This approach requires no prior knowledge about the objects in the image and no training data in order to find interest points in an image. This detector satisfactorily finds *Harris* corners from entropy coded images at each pixel neighborhood of the image. These corners mainly reside along the major object in the image. These sparse interest points are then tested for face recognition on standard databases. An improvement in the recognition results over the holistic representation has been reported by using sparse features. The details are given in Chapter 3.

- ***Action Recognition in HRI Applications:*** We study full body action and activity recognition by using a sparse spatiotemporal feature extraction approach, called *bag of words* (BoW) . This approach arises from content based image analysis and widely used over the past few years. A major benefit of this approach is that it does not require context information. Current outstanding challenges in BoW approach are the cluttered background and vocabulary formation. We extend this work at descriptor level by adding spatial features and achieve better results for facial expression recognition as compared

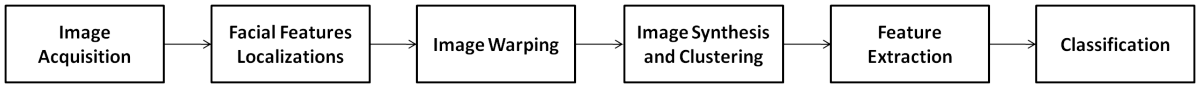


FIGURE 1.4 Sequential flow toward facial feature extraction and image preprocessing.

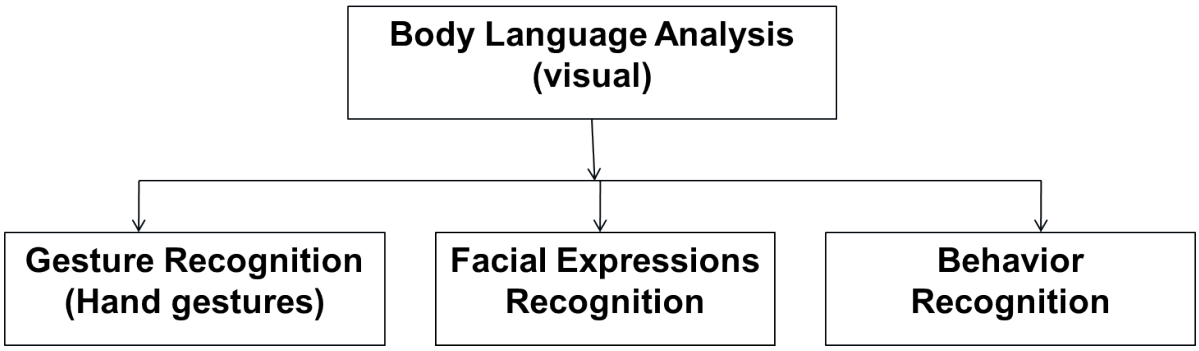


FIGURE 1.5 A generic vision based approach to full body analysis by [WRB10]

to conventional BoW. Moreover, We use varying cuboid sizes around the interest points and show from our experimental results that fixed size cuboid may cause loss of some information.

1.4 Thesis Outline

Human body language analysis has been one of the challenging tasks for the researchers in the area of computer science, psychology and robotics. Due to their universality, applicability and non-intrusiveness, human faces are widely used as compared to other biometric traits. The core of this thesis consists of Chapter 4 and Chapter 5 which represents face image analysis, whereas Chapter 6 deals with action recognition. In general, full body analysis includes study of three major traits called face image analysis, behavior analysis and hand gestures as shown in Figure 1.5 [WRB10].

Facial image analysis is performed using model based approaches which provide detailed description of the faces. Full body analysis is described by using a sparse spatiotemporal feature extraction approach. This problem is studied for human robot interaction (HRI) applications.

Chapter 1 discusses a general framework of this work. This includes motivation behind this research work, a detail description of the problem and the proposed solutions. The proposed solutions and contributions are described in detail in section 1.3.

Chapter 2 provides a detail and thorough literature survey. Related work is divided separately for different systems introduced in this thesis. We divide related work in two major parts, 1) face image analysis and 2) full body action recognition. Face image analysis is further divided in three parts which describes conventional approaches for face recognition, 2D face modeling and 3D face modeling. For full body action recognition we mainly focus on current content based approaches.

Chapter 3 gives a compact and self-contained implementation of a real time face image analysis system. Face image analysis consists of facial feature localization, face detection, image registration, face recognition and facial expressions recognition. This chapter also serves as a tutorial for implementing a face recognition system. It describes facial feature localization techniques, image normalization, image warping, image synthesis and finally feature extractions. Feature extraction is performed using different approaches. Further, this chapter introduces our proposed approach of interest point detection from entropy coded images. As a case study, we use this detector for face recognition. This chapter also explains facial expressions recognition using the same framework. This consists of eyebrow motions for controlling a humanoid robot.

Chapter 4 introduces the use of model based techniques for face image analysis. It describes various 2D face models and a detailed implementation of *Active Appearance Models* (AAMs) for STMF extraction is given. It starts from a basic definition of the models in computer vision and a compact representation of an object using model parameters. An overview of active contours, *Active Shape Models* (ASMs) and AAM is also given. Different model fitting modules are also discussed in detail. Finally, this chapter describes our point distribution model and its implementation to extract a useful set of features. We use this extracted feature set for face recognition, facial expressions recognition, gender classification and age estimation. This feature set is robust against facial expressions for face recognition applications.

Chapter 5 describes 3D face modeling using a single image. It proposes an improved version of STMF. The improvements are made in structural and especially in textural parameters. It describes different 3D face models and a detail implementation of our wireframe model. Shape of this model is defined differently as compared to ASM in Chapter 4. Furthermore, this chapter introduces texture mapping approach. This texture map is rectified from perspective distortion and is the useful tool against pose invariance. Feature set is extracted in similar way using structure, texture and temporal parameters. We use different textural features comparatively to study the effect of texture representation using texture mapping approach. The extracted STMF is robust against facial expressions and facial poses for face recognition applications.

Chapter 6 describes the action and activity recognition. It explains *bag of words* (BoW) approach for videos to extract spatiotemporal features. These features are extracted from sparse interest points which are extracted by applying spatial and temporal filters on videos. This markerless approach is applicable to videos and continuous image sequences. On facial expressions recognition application, we investigate that additional spatial features improve the classification rate as compared to conventional BoW.

Chapter 7 concludes this research work and provides some comprehensive conclusions. we explicitly describe a brief summary, conclusions drawn from this research work and some future extensions of this work.

CHAPTER 2

Related Work

We explain related work and literature survey separately for face image analysis and full body action recognition. The goal of this thesis is to extract the feature vector for body language analysis. We study different approaches for feature extraction. In general, a feature vector is a set of attributes which describes compact representation of the object or the image. A feature vector have following major characteristics [Ria04]:

- **Discrimination:** Features should take on significantly different values for objects belonging to different classes.
- **Reliability:** Features should take on similar value for all objects of the same class.
- **Independence:** The various features used should be uncorrelated with each other.
- **Dimensionality:** The complexity of pattern recognition system increases rapidly with the dimensionality of the system.

Section 2.1 addresses literature related to face image analysis. This section is further divided in two major sections. Section 2.1.2 explains literature related to conventional and recent image based approaches and toward the development of a face recognition system. Section 2.1.3 describes recent research work in model based face image analysis. We explicitly describe state-of-the-art 2D and 3D face models, their representation, development and applications. In section 2.2 full body action recognition with special focus on sparse spatiotemporal features is discussed. Further, in section 2.3 we provide classifier specifications used during the experimental evaluation of the proposed work. In the last section 2.4 we describe different databases comparatively along with their diversities. We study both face and action classification databases.

2.1 Face Image Analysis

In this section we describe in detail the methods and techniques for face recognition, facial expressions recognition and gender classification. We study context and content aware image based approaches, 2D models and 3D models in detail.

2.1.1 Human Faces in Cognitive Science

A formal study of the facial expressions was performed by Darwin in his book *The Expression of the Emotions in Man and Animals* published in 1873 [Dar73]. Recently, Toole [O'T09] describe three computational challenges in face recognition problem. These include:

- **Computational challenge I:** Human visual system acquire two 2D images of a face and reconstructs to acquire third lost dimension. This reconstruction can be error prone in certain cases. For example, a given facial structure and a varying facial texture may re-create several faces by just changing the lighting conditions on the texture.
- **Computational challenge II:** The neural system should encode and quantify the complex information in the faces. This includes uniqueness of individuals, properties that specify ages, sex, race and the social and emotion communication.
- **Computational challenge III:** The unique information that specifies face identity does not exist in absolute terms but rather depends on a reference population of relevant faces.

These challenges have been addressed by using human visual system, computer vision systems and their combinations. Toole [O'T09] claim that under certain scenarios human visual system may cause false alarms. This situation arises specially with previously unknown faces. A fusion between similar strategies results in minimal whereas distinct strategies (algorithm and human response) results in substantially improved performance. Sinha et al. [SBOR06] provide *nineteen results all computer vision researchers should know about* for processing of the face images by human visual system. An STMF finds its basis on some of these facts. These facts are categorized in five aspects 1) Recognition as a function of available spatial resolution, 2) The nature of processing: Piecemeal versus holistic, 3) The nature of cues used: Pigmentation, shape and motion, 4) Developmental progression and 5) Neural underpinnings.

David Matsumoto et al. [MW09] suggest that the facial expressions are independent of cultural learning and rather are innate. They are gene characteristics which are transfered by birth. According to Matsumoto, the statistical correlation between the facial expressions of

sighted and blind individuals was almost perfect. This suggests something genetically resident within us is the source of facial expressions of emotion [MW09].

A specific part of the brain, known as the *Fusiform Face Area* (FFA), is believed to process face stimuli more than other visual objects. The evidences also suggest that, as part of the human visual system, the FFA also processes categorical data of objects other than faces. A particular brain disorder can cause *Prosopagnosia*, that is face blindness, a face recognition disorder preventing people from identifying faces. *Prosopagnosia* refers to faces only, as all other visual stimuli undergo normal processing. An estimated two percent of all people suffer from face blindness. Also, because it is usually easier to recognize faces of races familiar to us, we may have already experienced some face blindness. For instance, if our environment comprises mainly Caucasian faces, we may have difficulty identifying Asian faces. This is known as the *Other-Race Effect*. Another face recognition effect is known as the *Inversion Effect*. Faces are unique in that when they are presented upside down, they are very difficult to recognize. The *Margaret Thatcher Effect* is another phenomenon which illustrates how difficult it is to identify distorted features on an inverted face, while these distortions can be easily identified on upright faces.

2.1.2 Image Based Approaches

Face recognition has been one of the widely studied topics over the last few decades. Due to the development of novel algorithms, availability of better hardware and their capability to deal with the real world scenarios, many face recognition algorithms are successfully performing in different real world applications. After the failure of *Bertillon system* [Ber09] for face recognition application, researchers started to pay attention to develop a reliable system to recognize the humans from their faces. This followed by several year research and efforts, resulting in several commercially available systems for face recognition [ZCPR03]. Generally, a face recognition system can be divided in three types:

1. Face recognition or identification is a *1:N* match where a given face is compared with N available face in the database.

2. Face verification or authentication is a *1:1* match where a given identity is verified against the claimed identity.

3. Watch list check is a problem where it is checked that either a person is available on the check list or not.

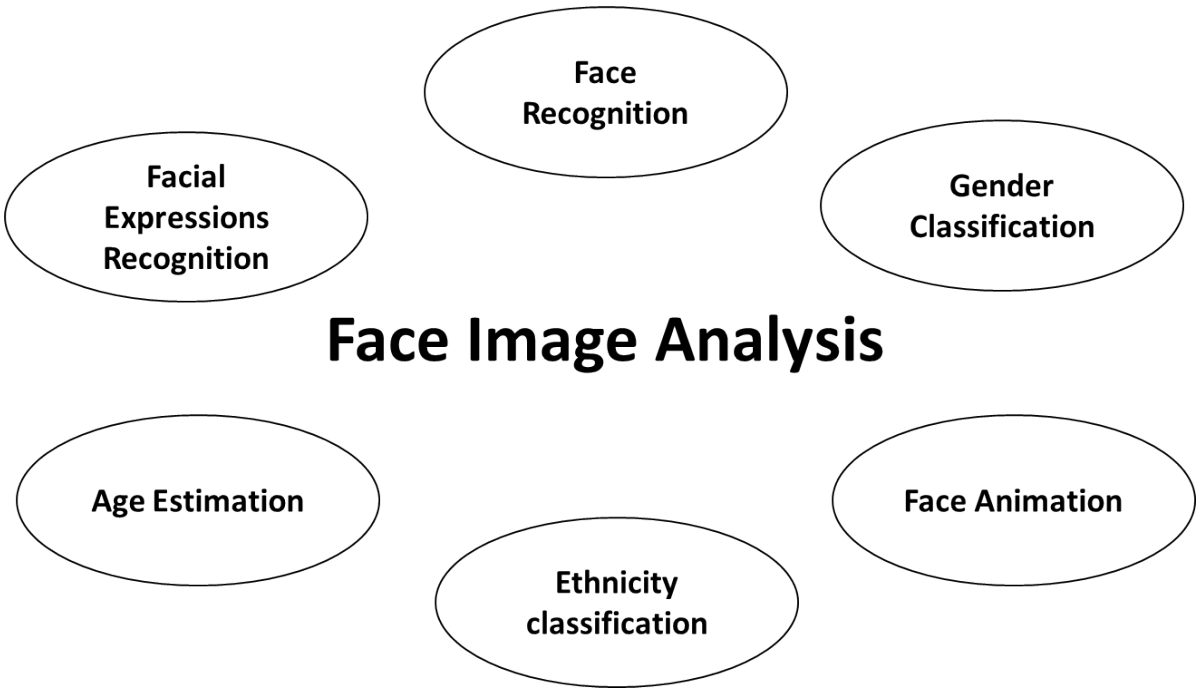


FIGURE 2.1 Face image analysis consists of face recognition, facial expressions, gender classification, age estimation, ethnicity classification and facial animation.

FIGURE 2.2 Generic flow of a face recognition system

Generally, face recognition systems consist of three major components (1) image preprocessing, (2) feature extraction and (3) feature classification. These modules are interrelated and occur in a sequence as shown in Figure 2.2. In [ZCPR03] authors describe different approaches toward the development of a face recognition system. They divide those approaches in three parts (1) feature based approaches, (2) template based approaches and (3) holistic approaches. In Chapter 3, we study feature based and holistic approaches in detail.

2.1.2.1 Holistic Approaches

Subspace learning from a training data of face images is generally treated as a linear problem. Faces are represented globally in a low dimensional subspace with a point. They do not necessarily address local facial features. For example, *Principal Components Analysis* (PCA) is

one of the extensively and successfully used approach for face recognition with its different variants [ZC05]. It is a global representation of the image and sensitive against slight pose and lighting variations. Initially it was used by Turk and Pentland in 1991 introducing the famous idea of *eigenfaces*. This is one of the highly cited work in the literature of face recognition [TP91a][TP91b]. However their work was inspired from Sirovich and Kirby's [SK87] approach of reconstructing the face image with face pictures from different weights and small collection of images. Though widely used, eigenface approach has severe issues with varying poses and illuminations. Several modifications have been proposed to this approach [ZC05].

Linear discriminant analysis (LDA), another holistic approach introduced in 1997 for face recognition by Belheumer [BHK97] which overcomes the issues of lighting variation. Each face is treated as a class while within class and between class variations are calculated. They treat face as *Lambertian* surface. The resulting face space maximizes the ratio of between class variance to within class variance.

Independent component analysis (ICA) has been successfully used for blind source separation problem [HO00]. Liu et al. [LW99] use ICA for face recognition and discuss its sensitivity of using with *Karhunen Loeve Transform* (KLT), *Fisher Discriminant Analysis* (FLD), *Maximum A Posteriori* (MAP) rule and *Bayes* Classifier. Further it is applied by [BMS02], where authors use two different architectures to classify faces in FERET database. One of the these architectures treats images as random variables and pixels as outcome, whereas the second one use the reverse of this order. The individual and combination of these architectures produce better results as compared to PCA.

2.1.2.2 Features Based Approaches

Local descriptors are recommended by some of the researchers over the holistic approaches because of their strong representation and robustness against occlusions [WMY+08]. In holistic approaches facial appearance averages over the image area and do not provide local feature description and spatial information. However local descriptors extract the texture from local regions [WMY+08]. Local features include SIFT [Low99], SURF [BTG06], LBP [OPH96], learning-based [CYTS10], spatiotemporal [DRCB05] and several others. *Local Binary Pattern* (LBP) is a binary coding of the pixel level information and is available with different versions. A conventional LBP works with (P, R) pattern [OPH96][AMH+06a], where P defines the neighborhood pixels and R is the radius. For example, $(8, 1)$ defines the circle around a pixel with eight pixels in neighborhood at a radius of one. A uniform LBP corresponds to two switches between the bits. For example 11100111, 01110000, 11110111 shows exactly two switches from 0 to 1 or 1 to 0. Spatially enhanced histogram shows additional spatial informa-

tion. We use conventional LBP in our experiments. A *volumetric local binary pattern* (VLBP) executes in three planes around a point of interest. These orthogonal planes are XT, YT and XY planes [VGM08].

Discrete Cosine Transform (DCT) is another strong representation for face recognition application. DCT coefficients provide low frequency information and are normally used with their different variants [Kha03] for image compression. It is used by Hafed et al. [HL01] for face recognition because of its compression power and efficiency. The underneath idea is the compression performed at retinal level is about 100 times the original object. Recent approaches rely mostly on feature descriptors, face modeling, bags of features and feature invariants. Cao et al. [CYTS10] recently showed strength of learning based descriptor over conventional LBP and *Histograms of Oriented Gradients* (HOG). They proposed a system to use machine learning technique to learn feature set from initially extracted raw feature vectors. A face image is divided in nine components including left and right eyebrows, left and right eyes, lip corners and forehead. A *difference of Gaussian* (DoG) is computed with $\sigma_1 = 4.0$ and $\sigma_2 = 2.0$. The ring based raw features are extracted from small regions which show illumination invariance. Final feature set is extracted after applying PCA to learned features. The system performed satisfactorily on LFW database database. We relatively tested similar approach and found satisfactory results on PIE database for face recognition across pose, illumination and expressions.

An important property of features is the invariance against real world challenges. *Scale invariant features transform* (SIFT) provides local descriptors with the strength over rotation, illumination and scale invariance [Low03]. Further, SIFT are invariant to image noise, slight occlusions and small changes in view points. This property of SIFT features has been successfully used in object classification [Low99] and hence for face recognition [BLGT06][LMT+07]. One of the very useful and compact representation of an image is histogram. Histograms can be built with different approaches. Dalal et al. [DT05] used HOG as fast and better tool for pedestrian detection. HOG are used widely in various applications including face recognition [AMM+08].

Sarfraz et al. [SHR10] used *face-GLOH* for face recognition. *Gradient location-orientation histograms* (GLOH) have advantage over the other methods because they work against varying poses and do not require alignment of gallery and probe images. *Local energy based shape histograms* (LESH) are used for pose estimation by Sarfraz et al. [SH08] in face recognition scenarios. Authors define a pose similarity feature space to restrict the problem to two classes i.e. inter-pose and intra-pose.

Speeded-up robust features (SURF) [BTG06] claimed to be faster than SIFT have been suc-

cessfully used by the researchers for face recognition. Steingrube et al. [SHN] use SURF due to their efficiency over 128-dimensional SIFT feature. Recently content based image analysis has played a significant role in object classification. *Bag of features* or *bag of words* (BoW) are one of the examples of these approaches used for object recognition.

2.1.2.3 Hybrid Approaches

Hidden Markov Models (HMM) have also been used successfully for face recognition. Faces are modeled as a *Markov model* by using the fact that facial components always lie in a sequences [NIH98]. A face image is divided in forehead, eyes, nose, lips and chin. Each component is modeled as a state of HMM. Nefian et al. [NIH98] use *Karhunen-Loeve Transform* (KLT) coefficients with HMM to improve recognition rate and efficiency of a face recognition system. Further modifications of HMM use *Embedded Hidden Markov Models* (EHMM) [Nef02], *hierarchical Dirichlet process hidden Markov model* (HDP-HMM), *maximum entropy Markov models* (MEMM) [NIH00].

Elastic Bunch Graphs (EBGs) are used successfully for face recognition based on wavelet jets. *Elastic Bunch Graphs Matching* (EBGM) was used by Wiskott et al. [WFKvdM97] towards face recognition in large dataset when only one image of a face is given. Such systems are useful in document classification where a person is available with only one image. The approach extracts fiducial points using bunch graph technique and calculates wavelet jets on these points.

2.1.3 Model Based Approaches

Modeling of the objects in computer vision requires context information of the images however, it provides compact representation of the object for which they are designed. Further, it provides various information about the objects including their geometry, texture, motion patterns and deformation modes. Generally models can be divided in two different categories on the basis of their functionality, 1) rigid models and 2) non-rigid models or deformable models. Rigid models define a rigid shape and texture of the object which can not be varied due to local structural and textural constraints in their design phase, examples include [May07][WMSR07]. Rigid models can synthesize global variations of the objects like their translations, rotations and scaling. On the other hand, deformable models are the models which can be adapted to different local and global variations of the object, examples include [CTCG95][CET98b][RMW+09][Ahl01][BV03b][WH04] etc. On the basis of their the structure, the configuration of the models can be divided in *point distribution models* (PDM),

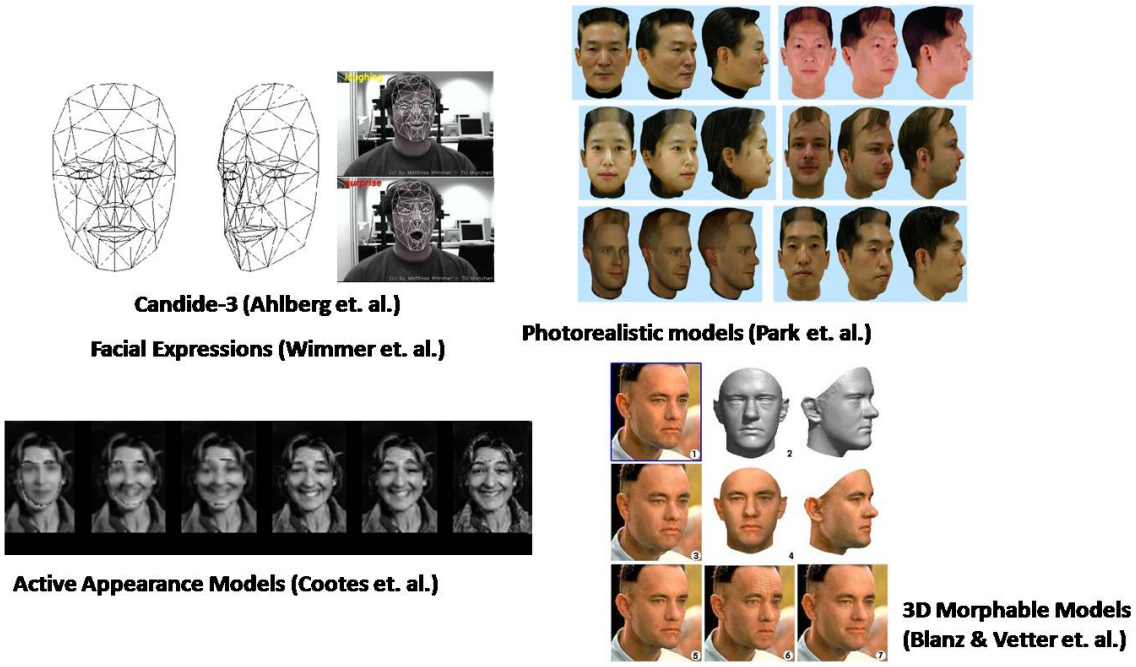


FIGURE 2.3 Example of different deformable models including, wireframe models, AAM, 3D deformable models and graphic photorealistic models. ©Images are taken from various sources.

point cloud distribution models acquired from laser scanners and wireframe models. Point distribution models are defined using a coarser distribution of the points in the space, examples include *Active Appearance Models* (AAMs) [CET01], *Active Shape Models* (ASMs) [CTCG95]. Wireframe models define points on the 3D surface of the object, examples include Candide-I,II and III model series [Ryd87][Wel91][Ahl01]. The projection of these models in 2D form a PDM and triangulated using the common methods like delaunay triangulations. Point cloud models are obtained from a dense 3D point cloud distribution. This distribution is generally obtained by using laser scanner, examples include *3D morphable models* [BV03b]. Other models are photorealistic models [PZVkC] and graphics models [Fac]. Figure 2.3 shows some of the deformable models studied in this work. Figure 2.4 shows different versions of the Candide model series. We study Candide-III and AAM in detail in thesis .

In the following section, we explicitly study the work related to 2D and 3D face modeling.

2.1.3.1 2D Face Modeling

Active contour models (ACM) also called *Active Snakes* are geometric and probabilistic models to define shape and its dynamics [KWT88]. *Active Shape Models* (ASMs) which are in-

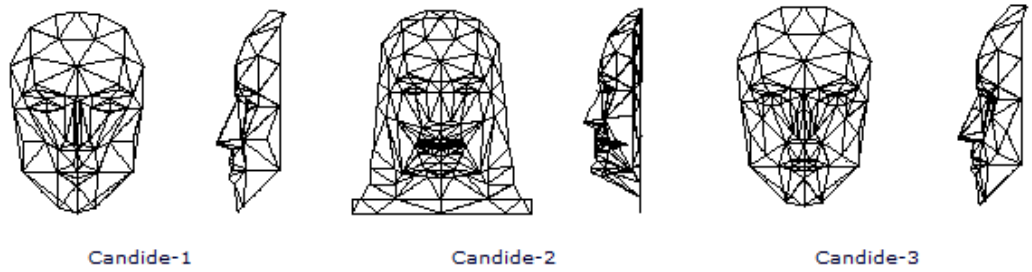


FIGURE 2.4 Different versions of the *Candide* face model. Candide-III is the latest version and used in this research work [Ahl00].

spired from active contour models are structural models consist of fiducial points defining location of different facial features. The baseline principle in their development is to achieve robustness in object analysis by keeping the control of geometric deformations caused by appearance and dynamical variations.

ASMs are point distribution shape models widely used in different computer vision applications problems [Ste04][WFS+08][RMWR08][CHC+94][WHS08]. Major challenges in ASM include robust image interpretation, model fitting to novel objects and real time capability. Since introduction of ASM in 1995, several modifications have been performed by the researchers. Model fitting is the adaptation of the model to the given image with optimal set of parameters which describes structure and appearance of an object. We study three different approaches for 2D modeling; 1)ASM fitting, 2) Objective functions and 3) inverse compositional image alignment, and use objective functions during our experiments. While for 3D face modeling we use an improved version of objective functions which uses displacements experts as a search domain at each vertex of the model [MR10]. ASM fitting relies on minimizing least square error and is slower and error prone. Wimmer et al. [WRMR08] use ASM for facial expressions recognition. They use objective function for model fitting. These objective functions are robust, real time and efficient to calculate. For 3D model fitting, we use similar methodology using *displacement experts* by Mayer et al. [MWS+08]. Matthews et al. [MB03] use *inverse compositional image alignment* (ICIA) method for model fitting and claim it much faster as compared to ASM fitting. Xiao et al. [XBMK04] extended ASM to 3D by using 2D+3D approach on the cost of more computations in calculating six additional transformation matrices than 2D models. Since model can be adapted to only those poses which are available in training set, hence it causes a misalignment if new poses are viewed. Vogler et al. [VLK+07] use best of two models for this solution. They combine ASM and 3D deformable models for model fitting in 3D and performing facial expressions recognition. Mil-

borrow et al. [MN08] use extended ASM for locating facial features in frontal views of upright faces. They further investigate several factors including effect of landmarks, adding noise to training images and using sparse covariance matrix on independent databases. Another result by Ramnath et al. [RBMR08] shows that dense model are better in fitting and performance. We verify this by using our PDM consisting of twice the number of points than that of conventionally used AAM. Amberg et al. [ABV09] study *forward compositional image alignment algorithm* for model fitting. Author propose two different methodology to overcome the issue of ICIA to fit a model to new faces, while maintaining the efficiency of the system. The idea behind is to use *Linearised Compositional descent* (LinCoDec) to increase the convergence radius of ICIA algorithm.

The type of models using both shape and texture parameters are called *Active Appearance Models* (AAMs) , introduced by Cootes and Edwards [CET98b]. These models are combined texture and shape models. Model parameters represent shape and texture implicitly. Since features set arise from different sources, hence they cannot be combined directly. Different approaches have been used in feature combinations [RGBR10]. A direct fusion may cause the dominance of the feature with high values and it also loses the orthonormality generated by different features. In order to generate combined model parameters, any suitable approach could be applied for instance, features normalization. AAM and 3D morphable models (details in section 2.1.3.2) are widely used due to their stronger representation [CET01][RMBR09b][MB03][ALC+09][Ste03]. In addition to texture and shape, we use temporal features for better representations of face image sequences. We experiment to show explicit role of shape, texture and temporal parameters on different applications using human faces. For example, performance of face recognition improves up to 13.65% [RMBR09a] by using additional temporal parameters (more details in Chapter 4 and Chapter 5). Further, a combination of three different features set lead to more flexibility in representation can be used for face recognition, facial expressions and gender [RMBR09c][RGBR09]. Such systems are very useful in human robot interaction scenarios, where the robots interacting with the humans should be intelligent enough to extract facial information in a stance.

In [ECT98] Edwards et al. use weighted distance classifier called *Mahalanobis* distance measure. However, Edwards et al. [ELTC96] isolated the sources of variation by maximizing the interclass variations using LDA. In [WRMR08] Wimmer et al. have utilized shape and temporal features collectively to form a feature vector for facial expressions recognition. Lanitis et al. [LTC95a] used separate information for shape and gray level texture. These models utilize the shape information based on a point distribution of various landmarks points marked on the face image. In our approach a predefined shape model consisting of 134 points is used.

Netzell et al. [NS08] use an approach to compute inner product of images which drastically reduces computational complexity of AAM fitting. In [HM09] Hamsici et al. use rotation invariant kernel for AAM fitting. They study 3D morphable models and AAM comparably in their approach. Lee et al. [LK09] propose tensor-based AAM for fitting with two different tensors called image tensor and model tensor.

2.1.3.2 3D Face Modeling

3D human face modeling has proved its capability to deal with a current outstanding challenges in face image analysis. 3D models have the capabilities to deal with varying view points of the objects and modeling the light variations. A detailed study of the deformation in the object structure is possible by controlling with only a few parameters [WH04]. On the other hands the system might be less efficient to be directly applied in real world [BV03b]. As compared to 2D modeling, 3D face modeling provides the lost information about depth, light source and structure of the object. This section focuses on the modeling of human face using a three dimensional model for shape model fitting, texture and temporal information extraction and then low dimensional parameters for recognition purposes.

3D morphable models are one of the recent state-of-the-art approach for modeling a realistic face [BV03a][Bra01][RV03][Rom05][HHB03]. The model is acquired by the 3D-texture scans instead of images or videos. Shape of the model is represented by a limited set of parameters $\rho = [f, \phi, \theta, \gamma, t_x, t_y, \mathbf{t}_w^T]$, where f is the focal length, ϕ, θ and γ are yaw, pitch and role respectively, $\mathbf{t}_w$ is 3D translation [ZC05]. This facilitates in modeling the face from scans instead of raw images. 3D morphable models showed their high potential against various facial poses and illumination modeling. Romdhani et al. [RV03] proposed a faster model fitting approach. A component based morphable model produced better results for face recognition as compared to conventional morphable models [HHB03]. Model synthesis is performed by inverse mapping from image to (u, v) texture plane. Bronstein et al. [BBK03] used geodesic distances for face recognition against facial expressions without reconstructing the 3D face. Park et al. [PJ07] apply 3D model for face recognition on videos from *CMU Face in Action* (FIA) database. They reconstruct a 3D model acquiring views from 2D model fitting to the images. Riaz et al. [RMBR09a] use spatiotemporal features from 3D model for face recognition against facial expressions. Author also describe the comparative performance with similar 2D model based features and improvements by temporal features.

Another important representation of the faces is the behavior of the persons, which is interpreted in the form of facial expressions. Other facial traits which do not directly represent person identity but play a significant role are categorized as *soft-biometrics* [JDN04]. These

include facial expressions, gender, age estimation and ethnicity [RMBR09d]. In our daily life activities, humans are capable in extracting this information in the very first interaction however concrete work for comparable performance in robotic vision is still missing in literature of face image analysis due to various outstanding challenges. In [RMBR09d] Riaz et al. use wireframe model for face recognition, facial expressions and gender classification. A model of an object is generally a set of parameters which describes the object in detail. Our proposed features are the parameters which fulfill a model's basic requirements like providing control over non-rigidness, deformation and capability to synthesize novel views. We use rather a coarser model called Candide-III [Ahl01] which is defined with 113 vertices and 184 triangular surfaces. This model is supported by action units and MPEG-4 animation units [LJ05] which makes it useful for facial expressions analysis. This model is a shape model but can be realized with texture.

We emphasize on 3D modeling for the faces because of the several benefits. One of the major advantages of model parameters is the better control over face dynamics, structure and appearance. Model parameters have been varied by Vetter et al. [BV03b] to control expressions and gender. Further, 3D morphable models are robust against pose and lighting variation. Our system solve similar problem but using very coarser model with normal camera images and achieve accuracy and robustness. In [RMBR09d] Riaz et al. propose a multifeature fusion but robustness of the system against facial poses and expressions is not studied. In [ASWG09] Asthana et al. have tested several techniques towards model fitting and their performance for facial expressions on CKFE database. We compare our results with their results but using different fitting techniques.

2.2 Full Body Human Activity

Full body action recognition has been one of the widely studied approaches over the last few years. We mainly study *bag of words* (BoW) approach thoroughly in this thesis. This approach is adapted by the computer vision researchers from *natural language processing* (NLP) and successfully used for information retrieval in documents. It is applied by learning vocabulary in the sentences without preserving grammar information. The major advantages of using BoW is their strong representation and formation of vocabulary for action classification. It has been successfully adapted in computer vision by Li [FFP05]. Despite their successful use in document analysis and computer vision, there are several challenges and built-in limitation of this approach. For instance, position of the words is not preserved, multiple objects in the images or cluttered images and vocabulary formation. Recently several modifications have been

performed by the researches in BoW [DRCB05][VGM08][MSH+10][RLS09][LYS09][LLS09]. The interest point detection does not correspond to anatomical features in the image. For example, in action recognition these interest points might not represent different body part but rather motion features. Similar motion features are collected in one bin. Major advantages of these approaches are that they do not require object detection, localization of body parts and segmentation.

bag of words (BoW) or *bag of feature* requires no expert annotation for learning and categorization and codebook is formed by using unsupervised learning [FFP05]. Feature selection and extraction to initialize BoW have been performed in different ways by the researchers. These methods include regular grids [Fei05][VS02], interest point detection [KB01a][VnU03], motion features [DRCB05][LMS+] and other methods including random sampling [MGPW05] and segmentation [BDF+03]. Once features are extracted, they are used for codebook formation or descriptor extraction. Key point detectors used in these approaches are SIFT [Low03], Harris detector [HS88], HoG [DT05] and spatiotemporal detector [DRCB05]. Vocabulary formation is generally performed using clustering. *K-means* clustering is mostly used by researcher to form vocabularies. Other approaches used feature-trees [RLS09]. Dollar et al. use k-means clustering after performing PCA, where K is taken as the number of eigenvectors retained. Further vector quantization is performed. For classification purpose, generally support vector machine (SVM) is applied with leave one out approach.

Different solutions have been proposed by the researchers to incorporate spatial relation between the feature vectors. Matikainen et al. [MHS10] use pairwise spatial and temporal representation for action recognition. An augmented use of *space-time interest points* (STIP) [LMS+] using *histograms of oriented gradients* (HOG) and trajectory based feature [MHS09] [MPK09]. This fusion has been proved to be reliable against noise and computationally efficient. Laptev et al. [LMS+] use scripts to annotate different videos clips acquired from Hollywood movies. Scovanner et al. [SAS07] extend SIFT to 3D domain for action recognition applications in videos using bag of words approach. Since local and holistic descriptors show different representations of the actions classes, hence Sun et al. [SCH09] use a fusion of local and holistic features for action recognition on *kth* and *Weizmann* database. Liu et al. [Sha08] use multi-feature for action recognition. Liu et al. use spatio-temporal cuboids and spin-images, which aims to capture the shape deformation of the actor by considering actions as 3D objects (x, y, t). Kellokempu [KZP08] use local binary pattern for dynamic texture extraction for action recognition applications.

2.3 Classifiers and Classification Criteria

In general, the criteria for classifier training and testing is similar throughout the experimentation in this thesis. we used three approaches for training and testing of the classifier.

- We use *k-fold* cross validation for classification. In most of the experiments on face recognition, facial expressions and gender classification, we use *Decision Trees* (DT) and *Bayesian Network* (BN). Cross validation is one of the best approximations for classification when dataset is limited and provides the unbiased behavior of data. In this approach, whole data (extracted feature vectors) is divided into *k* independent subsets and training of the classifier is performed with *(k-1)* parts while the remaining one subset is used for testing. This process is repeated *k*-times and results are averaged over *k* outcomes. In our experiments, we take $k = 10$ which is also used in common practices by researchers. This approach also overcomes over fitting problem.

- In addition to *k*-fold cross validation, we further use classifiers with same parameters configuration of the classifier for training and testing using data split. This split is in such a way that two third of the data is used for classifier training and remaining data is used for testing purpose.

- For smaller dataset like, our lab captured images and FG-NET database dataset where very less number of subjects are available in the database, we use leave one out criteria. In this classification process, we use one class for testing while training the classifier with the remaining class data. This process is repeated as much number of times as the number of classes in the data.

Leave-one-out cross validation is a special case of *k-fold* cross validation, where *k* = number of classes in the data.

Choice of the classifier is made on the basis of the given applications. We prefer to use decision tree for classification because of the several reasons. Our feature set consists of multiple features and it is recommended to use decision tree since it goes through each attribute of every observation until the leaf node is purified. This makes them more suitable to classify at low levels. On contrary Bayesian network is a graph which defines a relationship between the attributes. These graphs are less detailed than the decision tree and hence their performance is relatively low as compared to decision trees. However, the choice of classifier in our results is not final and other classifiers may perform better depending upon the structure of the classification problem. For example, in order to verify our results we also have tested with support vector machines and random forests. These classifiers also perform satisfactorily.

Method	Pros	Cons
k-fold cross validation	Accurate performance estimations, avoids over fitting	Overlapping training data, Small samples of performance estimation
Leave-one-out cross validation	Unbiased performance, good for continuous error estimation	Very large variance
Split data	Fair classification in a given database	Random selection for the split, repetition in experiments

TABLE 2.1 Comparison of three validation methodologies used in our experiments.

Classifier Spec.	Value
Confidence Factor	0.25
Min- number of instances per leaf	2
number of folds	3
Tree pruning	true
Pruning Approach	C.4.5
Classification	10-fold cross validation

TABLE 2.2 Parameter specification used for decision tree classification. These parameters are kept constant throughout the experiments.

For Bayesian network we use specifications given in Table 3.2.

2.3.1 Decision Trees

Decision trees are non-metric classifier for numerical and nominal data. A tree is initialized with a node which is called root node. A decision is made at each level with some given splitting criteria. This splitting is called factor *B* and depends upon the type of the tree. For a binary decision tree, $B = 2$. The splitting criteria might be entropy impurity, variance impurity, Gini impurity or classification impurity. This test at each node assures the purity level for the decedent node. Instead of maximizing the purity, impurity can also be reduced, ideally equal to zero. The tree grows unless a stopping criteria is defined. For stopping the tree different termination criteria are used e.g. cross-validation, thresholding, *minimum description*

Classifier Spec.	Value
Estimator	SimpleEstimator
search Algorithm	K2
ADTree usage	false
Estimator	SimpleEstimator -A 0.5
alpha value (initial estimate)	0.5
searchAlgorithm	K2 with Bayesian Network

TABLE 2.3 Parameter specification used for Bayesian network classification. These parameters are kept constant throughout the experiments.

length (MDL), hypothesis testing or using confidence factor on some given distance metric. Further, pruning can be used as a tool to correct for potential over fitting. There are different types of trees depending upon the applications. The famous are ID3, C4.5, J48 and BDT.

We use J48 decision tree from Weka in our experiments which uses C.4.5 algorithm. Specification which we use during the experiments are shown in Table 3.1. J48 gives us the best recognition rate and is recommended to be used in case of multiple features. We prefer to use tree-based classifier on Bayesian networks. Experimental evidence of their use is given in Chapter 4 and Chapter 5.

2.3.2 Bayesian Network (BN)

A Bayesian network is a graphical representation of the data using probabilistic relations among the set of variables. If $U = \{x_1, \ldots, x_k\}$ where $k \geq 1$ be a set of variables. A Bayesian network B over a set of variables U is a network structure B_S which is a directed acyclic graph over U and a set of probability table $B_P = \left\{ p(u|pa(u))|u\epsilon U \right\}$ where $pa(u)$ is the set of parents of u in B_S. A Bayesian network represents a probability distribution $P(U) = \prod_{u\epsilon U} p(u|pa(u))$. For details refer [Bou05a].

We use simple estimator for BN. It produces direct estimates of the conditional probabilities by using following relation:

$$P(x_i = k|pa(x_i = j)) = \frac{N_{ijk} + N'_{ijk}}{N_{ij} + N'_{ij}} \tag{2.1}$$

where N'_{ijk} is the alpha parameter which is set to 0.5 throughout our experiments. A value $alpha = 0$ approaches to maximum likelihood estimates. Search algorithm used hill climbing

approach K2 [CH92]. Table 3.2 represents specifications for Bayesian network used during the experimentation. For further details, refer to [Bou05b][Hec95].

2.3.3 Support Vector Machines (SVM)

SVM is a binary classifier and used to separate two classes with a line between them. If the given observations are $\{x_i, y_i\}$, where $i = 1, \ldots, n$ and n is the total number of observations, $x_i \epsilon R^d$, where d is the dimensions of the space and $y_i \epsilon \{-1, +1\}$ are the two labels for the given data. The purpose of an SVM classifier is to find a line $y_i = mx_i + c$ with some slope m and y-intersection c such that the values with $y = +1$ lies on one side of the line $y_i(x_i) > 0$ and the values with $y = -1$ lies on the other side of the line $y_i(x_i) < 0$. However if we have a multi-class problem then the aim of an SVM classifier is to find hyperplanes which describes the data with the best separation. We use LibSVM [CL01] for experimentation.

2.4 Databases

In order to validate the representation power of our proposed feature set we experiment on various standard databases. The experiments are performed on those databases which have a close relation to real world application. For the face image analysis, we use benchmark databases, which provide sufficient diversity in varying poses, facial expressions, non-uniform lightings, aging effect and different ethnicities. We use Cohn Kanade Facial Expressions (CKFE) database [KCT00], Man Machine Interaction (MMI) database [MSV+09], CMU Pose Illumination and Expressions (PIE) database [TBB02], AT and T database [orl], Yale database [GBK01] and FG-NET database for facial age estimation [fgn].

Most of the experiments have been performed on Cohn Kanade Facial Expressions (CKFE) database for human faces. Some examples of CKFE database are shown in Figure 2.6. We experiment on frontal views of the database. The CKFE database contains 488 short image sequences of 97 different persons aging between 18 to 30 years, performing the six universal facial expressions. It provides researchers with a large dataset for experimenting and benchmarking purpose. Each sequence shows a neutral face at the beginning and then develops into the peak of an expression. Furthermore, a set of *action units* (AUs) has been manually specified by licensed *Facial Actions Coding System* (FACS) experts for each sequence.

CMU Pose Illumination and Expressions (PIE) database was captured between October and December 2000 consisting of 41,368 images of 68 people. Each person is captured in CMU 3D Room under 13 different poses, 43 different illumination conditions, and with 4 differ-

FIGURE 2.5 Example images showing pose variations in CMU Pose Illumination and Expressions (PIE) database .

ent expressions. Furthermore, the captured faces are from various ethnic origins [TBB02]. Figure 2.5 shows pose, expressions and illumination variations of this database.

The AT and T database (formally called ORL database) consists of ten different images of each of 40 distinct subjects. For some subjects, the images are taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position [orl]. This database is suitable for the applications where less example images per person are available.

Yale database contains 165 grayscale images in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink [GBK01].

The Face and Gesture Network (FGNet) database is an image database containing face images showing a number of subjects at different ages. The database consists of images from 82 persons with overall 1002 images. The age range from 0-69 years [fgn].

Man Machine Interaction (MMI) database is designed for facial expressions recognition and

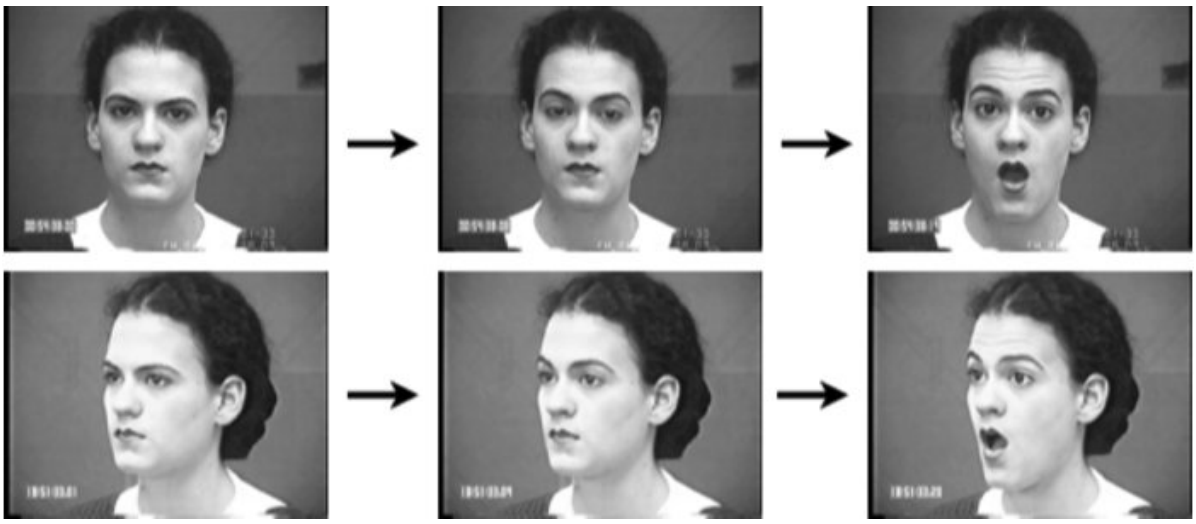


FIGURE 2.6 Examples from Cohn Kanade Facial Expressions (CKFE) database .

FIGURE 2.7 Examples from our lab captured images for five different facial expressions (from left to right) anger, disgust, surprise, sadness and pain (*courtesy: University of Hannover, Germany).*

consists of six basic facial expressions [MSV+09]. There are 2894 sessions in this database out of them 1395 sessions are AU coded and 197 sessions are labeled as one of the six basic emotions.

In addition to the aforementioned databases, we also experiment on laboratory captured images sequences and live demonstration in different workshops and exhibition in indoor environment [Ria08][Ria09][Ria10]. Our laboratory captured images are also acquired from Pan-Tilt-Zoom (PTZ) camera in an assistive environment. Face recognition and facial expressions recognition is performed together with fall detection for elderlies. We benefit from zoom feature of the camera and run face image analysis module whenever a face is frontal to the camera. Some examples from our laboratory captured database are given in Figure 2.7. We add an additional *pain* expression, which is generally not available in standard facial expression databases. This expression is artificially generated by using FACS [EFH02] coding.

FIGURE 2.8 Examples of a pick and place action observed from a ceiling camera. The images shows every 20^{th} frame from a given sequence [TBB09].

For full body action recognition, we use TUM-kitchen database [TBB09]. Other widely used databases for action recognition include Weizmann-database [GBS+07], kth-database [SLC04] and UCF-sports database [RS10].

TUM-kitchen database consists of detailed low level actions and designed to study human robot interaction, human motion tracking, motion segmentation, and activity recognition. Different subjects perform table setting activity by grasping, picking and transporting the objects from one place to the other. These tasks are person dependent for instance a person might pick up the object from left or from right hand. Further persons might also pick multiple objects at a time. Some of the activities are annotated by humans where they perform activities like a robot [TBB09]. Some examples of this database are shown in Figure 2.8.

kth-database is one of the widely used database for action recognition experiments. It consists of six different types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. The database contains 2391 sequences and all sequences are taken against homogeneous backgrounds with a static camera [SLC04]. Weizmann-database consists of 90 low-resolution (180 x 144) video sequences of nine person performing 10 natural actions. These actions contains run, walk, skip, jumping-jack, jump-forward-on-two-legs, jump-in-place-on-two-legs, gallop sideways, wave-two-hands, wave-one-hand and bend [GBS+07]. UCF-sports dataset consists of video sequences taken from youtube. The primary focus is to provide the computer vision community with an action recognition dataset consisting of realistic videos. The challenges in this database contain large variations in camera motion, object appearance and pose, object scale,

viewpoint, cluttered background and illumination conditions. There are 50 categories in this database, the videos are grouped into 25 groups, where each group consists of more than 4 action clips. The video clips in the same group may share some common features, such as the same person, similar background, similar viewpoint, and so on. For more details, we refer to [RS10].

CHAPTER 3

Face Recognition for HRI Scenarios

This chapter addresses a real time face recognition system capable to perform in daily life activities. The approach followed in this chapter includes local facial feature detection, image registration, face recognition, facial expressions synthesis on a humanoid robot and sparse feature extraction for face recognition system. This chapter also serves as a tutorial to develop a standalone system for real time face image analysis. Besides the development of the recognition system, this chapter also provides some solutions to deal with some of the current outstanding challenges in face recognition. These challenges include varying poses, illuminations and lightings, facial expressions, partial occlusions (which also include makeups and facial hair) and aging effects. The system introduced in this chapter is adapted from conventional face recognition approaches and robust against limited in-plane and out-of-plane rotations in the presence of facial expressions. Similar approach has been adopted by different other researchers for image normalization [Eke09]. After thoroughly preprocessing the images, various feature extraction approaches are applied. In order to verify the performance of the system, experiments are conducted on standard databases. In addition to this, we study an interest point detector from entropy coded images which is used for face recognition as a test case. The entropy coded images are obtained by replacing each pixel value with its entropy value in the neighborhood. Local spatial features are extracted from these interest points to form a feature vector.

3.1 Problem Statement and Solution

In human robot interactive scenarios, faces are seen in action, exhibiting different meaningful deformations and dynamics which convey a large amount of information. These factors degrade facial recognition results. Conventional holistic approaches are quite efficient in face recognition in constrained environments however they produce unsatisfactory results for real

world applications. To overcome this problem, a solution devised by the researchers is to preprocess face images for representative feature extraction. The benefits of this approach are two fold, 1) It utilizes the efficiency characteristics of holistic approaches, 2) It makes the system robust against different variations. However, the robustness of this system is still limited which urges to construct a detailed model of the faces for better representation. Face modeling is robust in real world challenges however less efficient as compared to image based approaches. Figure 3.1 shows our approach followed in this thesis to develop a face recognition system in real time. In order to deal with slight facial expressions and limited in plane and out of plane rotations, we use image normalization techniques which are simple and used by different researcher [Eke09]. We use both eyes positions, nose and lip positions which are robust to lighting and head rotations. These landmarks are used to normalize the given face images on a standard template using image warping. These preprocessed images are then used for face recognition and classifying facial expression in face-to-face interaction with humanoid robots.

3.2 Image Registration

In general practices, images captured from camera require normalization before feature extraction. In our system, any input face image is registered using three points based affine transformation. These three points include both eyes and lip positions. Firstly, we use face detector to find either there is face in the image or not. We used *Viola and Jones* face detector for this purpose. The details are given in section 3.2.1. Once a face is found in the image, eyes are searched in the face window. The eyes detection search can be limited to the region of interest inside face window and a constraint is applied on the vertical distance between the positions of both detected eyes (see equation 3.1). Eyes detection method works robustly during the whole recognition phase. Similarly lips detector search is also performed in the detected face window area by using lip color. If face is not found, local features cannot be located. If face detector falsely detects a face then facial features are also detected falsely. In order to prevent this we apply a further constraint on local features. This constraint restricts distances between the position of the two eyes. If E_l and E_r are the positions of left and right eyes respectively, then we apply a constraint on vertical distances, that is

$$|E_{ly} - E_{ry}| < \theta \tag{3.1}$$

Where θ is the predefined threshold which is set roughly to six pixels in our experiments. Where E_{ly} and E_{ry} are the vertical positions of the left and right eye respectively in the given image.

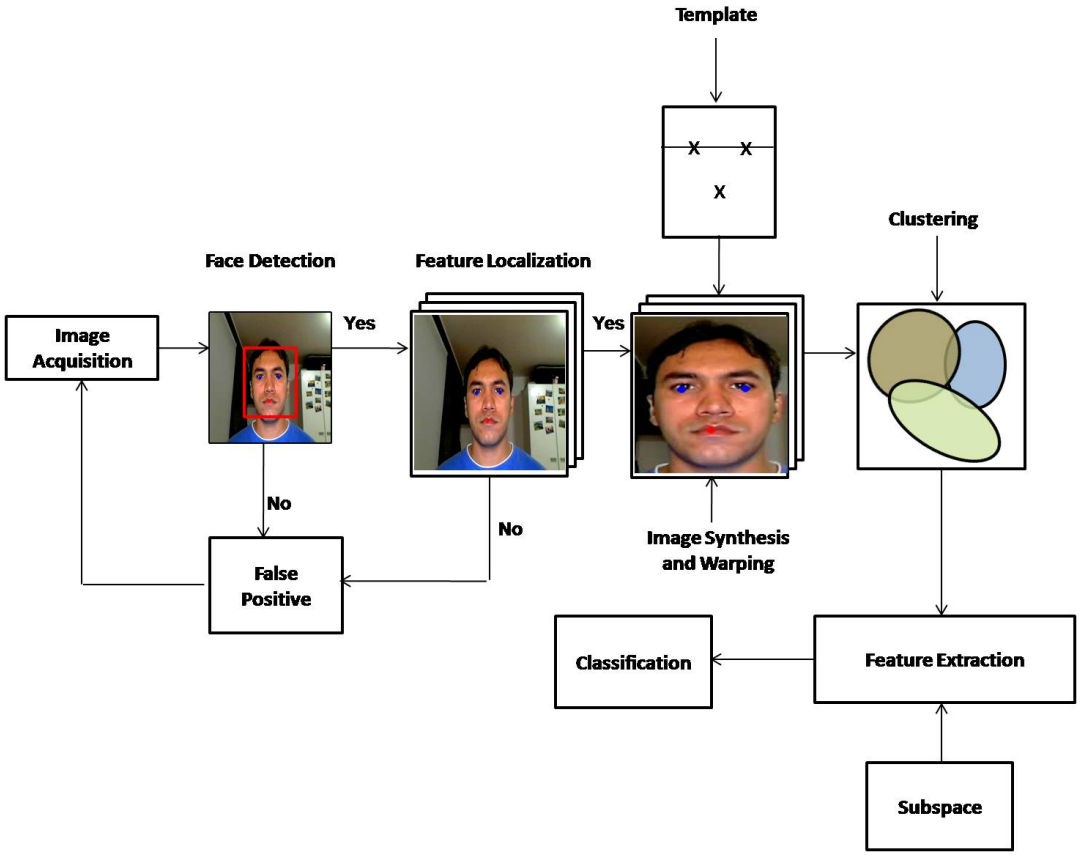


FIGURE 3.1 Face recognition process: The system acquires image through a camera or from a memory location, detects face in the given image. If face is not found the system searches the next image for face detection. In detection window facial features are detected. The process is repeated again with the next face image if facial features are not detected in the current image. The successful image is normalized and synthesized with four additional lip positions (x_i, y_i), where i = -1,0,1, in the neighborhood of the detected lips (x_i, y_i). Different features are extracted from these images.

3.2.1 Face Detection

For initialization of our algorithm, we use face detection from Viola and Jones [VJ04]. This real-time object detection module is widely used by research community and embedded in most of the computer vision libraries and toolkits. We use *opencv*-1.0 implementation which is capable to work in real time for such scenarios. The algorithm is based on three major parts:

- Calculating Haar-like features, which are shown in Figure 3.2 from integral images which are calculated from equation 3.2.
- Classifier training and choosing the features from large set of positive and negative

example images.

- Creating a cascade of weak classifiers using AdaBoosting.

An idea of integral image is introduced by [VJ04] for fast calculation. The value of the integral image ii at any point (x, y) is the sum of all previous horizontal and vertical pixel values. It is calculated by:

$$ii(x, x) = \sum_{x' \leq x, y' \leq y} i(x', y') \tag{3.2}$$

Where $ii(x, y)$ is the integral image and $i(x, y)$ is the original image and $0 \leq x' < x$ and $0 \leq y' < y$.

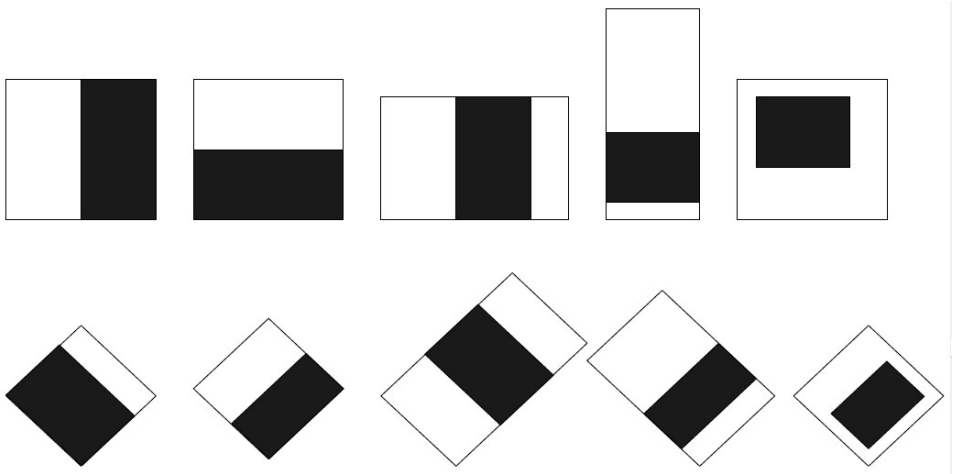


FIGURE 3.2 Haar-like features of different sizes and orientations. The sum of pixel values in white region is subtracted from the sum of the pixel in black region.

The performance of the detector depends on the set of training images. The classifier normally does not work for high rotations and bad lighting conditions. Further the occluded and blurred faces are also hard to identify. However the performance of the detector in cluttered backgrounds is quite satisfactory. Figure 3.3 shows some examples of true positives and true negatives of the face detector used in our case.

3.2.2 Facial Feature Detection

We evaluate different detection approaches for facial feature localization. We use haar-like features [VJ04] for face detection, template based matching [Bru09] for eyes detection and

FIGURE 3.3 Face detection results: The detection algorithm works fine with frontal and slight profile faces but fails for rotated face. In above examples it works for out of plane rotations but fails for in plane rotations.

color based detection [MWR09] for lip detection. In the following section, we explain in detail about different facial feature detectors.

3.2.3 Eyes Detection

Haar-like features are strong representation for object detection. These features are extracted from a large database of positive and negative example images to train a classifier for object detection. Haar-like features were introduced for object detection [VJ04]. These features are fast to calculate from integral images. However, they require training a cascade from a large set of training images. We use rather a simple approach called template based matching which requires no initial training. We make two separate templates for left and right eye and match it within detected face window using correlation based template matching. If $I(x, y)$ denotes the test image, $T(u, v)$ is the template to be matched the correlation coefficient $c(u, v)$ is given by:

$$c(u, v) = \sum_{x,y}[I(x, y)T(x - u, y - v)] \tag{3.3}$$

The normalized correlation coefficient overcomes the difficulties of varying lighting conditions and correlation values between true and false templates by normalizing the image and feature vectors to unit length.

$$\gamma(u, v) = \frac{\sum_{x,y}[I(x, y) - \overline{I}_{u,v}][T(x - u, y - v) - \overline{T}]}{\sqrt{\sum_{x,y}[I(x, y) - \overline{I}_{u,v}]^2 \sum_{x,y}[T(x - u, y - v) - \overline{T}]^2}} \tag{3.4}$$

Where $\gamma(u, v)$ is the normalized correlation coefficient [Lew95]. We use *opencv* implementation for template matching separately for left and right eye. A given template is searched along the whole face window and correlation is calculated. A maximum correlation indicates

the presence of that template in the query image. Our implementation uses *CV_TM_CCOEFF_NORMED* which utilizes normalized correlation coefficients. Eyes search is restricted to a horizontal axis. An image is discarded if both eyes are not located on a horizontal axis.

FIGURE 3.4 Eyes and Lip detection results. Eyes are detected using template matching and lips detection is performed using lip color.

3.2.4 Lips Detection

Lip position is determined using lip color mask in face window which is distinctive to the skin color. This is a pixel based approach which introduces adjusted pixel features for different facial components. An adjusted pixel feature is extracted from static pixel feature and image characteristics. For further detail refer [MWR09]. The algorithm is capable to handle diversity in skin and lip color due to varying ethnicity. Similar types of pixel values cluster together and benefit in classifying as different facial components. This approach uses a color mask extracted from training images. For any input image, the probability of each pixel is evaluated with the pre-calculated color mask and assigned to a specific facial component within a region of interest. Figure 3.5 shows an input image with facial features localization and normalized image.

3.2.5 Image Warping and Clustering

Image warping is performed using three detected facial features. We define a standard template for face images. All input images are warped to the standard template. This approximate template can easily be adapted from any frontal image and does not vary throughout the experimentation. Since eyes are located with distance constraint, so their position is found on a horizontal axis. However lip detector is relatively unstable because of lip motion and size. We

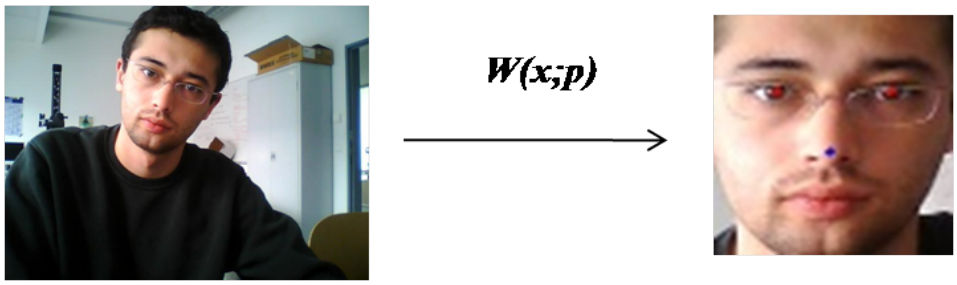


FIGURE 3.5 Image registration process. Facial features are detected such that the eyes lie inside face window on an approximately same vertical level. Using three points image warping, face image is normalized to standard template. This template is arbitrary and fixed throughout the experiments. $W(x;p)$ represents affine warping at position x with pixel value p.

synthesize each input image with four additional lip positions in the neighborhoods of the detected lip position. These additional positions include up, down, left and right neighborhoods of the detected position. In this way single input image is additionally synthesized to four different images with four different lip positions. Once these images are synthesized, we get four times more images. During experiments, we collect 60 images for each subject and synthesize with four different lips position in the neighborhood of the detected lip position. At this stage, $60 * 4 + 60 = 300$ images are obtained per single person. In order to reduce the data size, we use k-means clustering with *k = 60*. After k-means clustering, we obtain well aligned images which are suitable for feature extraction. This pre-processing enhances the performance of holistic features. Similar approach has also been used by [ESTS09].

Affine warp is a useful tool for image registration. The transformed images preservers the line parallelism properties and equispacing between the points. An individual transform or a combinations of transformations can be used to transform an images. These transformations include rotation, scaling, shearing, flipping and translation. Equation 3.5 shows affine transformation between two images, where x_o and y_o are the discrete output image locations, x_i and y_i are the inverse-mapped input locations, and $a_{11} \ldots a_{23}$ are the six affine warp coefficients.

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} x_o \\ y_o \\ 1 \end{bmatrix} \tag{3.5}$$

Where $0 \leq x_o < w - 1$ and $0 \leq y_o < h - 1$. Where w is the width and h is the height of the image. Further we fill all those pixel of target image to zero, if they correspond to outliers in

the source image. In order to solve equation 3.5 we need at least three points correspondence in target image and source image. These three points are left and right eye positions and lips position. We do not use nose because the lip detector is more robust and distinctive in nature due to its color.

3.3 Feature Extraction

We use three different features due to their strength of representation and efficiency. We use *Principal Components Analysis* (PCA) , *Discrete Cosine Transform* (DCT) and *Local Binary Pattern* (LBP) . Further, we also study sparse representation of the face images which shows improved performance in face recognition applications.

3.3.1 Principal Component Analysis

Principal Components Analysis (PCA) is the widely used and one of the very first approaches to face recognition in terms of *eigenfaces* [TP91a]. Images are treated holistically and extracted features do not correspond to local facial features. However, this approach is highly sensitive to facial poses and lighting conditions. Since images are pre-processed to tackle with misalignment issues, it is meaningful to apply PCA at this stage. However, PCA relies on training a subspace method. With every new person in the database, subspace learning is repeated on the whole database. Each image is resized to 64×64 after normalization. We collect n images where $n = 60$. If I_i be any ith image then all images are stored in a matrix D column wise for PCA analysis, where $D = [I_1, I_2, \ldots, I_n]$. The size of each column is $64 \times 64 = 4096$. In our face recognition system, we use 60 images per person for subspace learning. These images are obtained after preprocessing. We obtain 97% of covariance energy during subspace learning. Feature vector consists of coefficients of the eigenvectors.

3.3.2 Discrete Cosine Transform

On contrary to PCA, *Discrete Cosine Transform* (DCT) requires no subspace learning. DCT coefficients represent low frequencies which are useful for person specific information.

We use block based discrete transform coefficients as a feature vector of a single image. We use block based feature extraction approach which is also used by different researchers [HL01][ESTS09]. Each image is divided in small blocks of 8×8. Coefficients are extracted from each block in a zig-zag pattern. The is shown in Figure 3.6. We extract only five

coefficients from each block which represent high frequency information. For a given image *I*, if total number of blocks are N_b, then size of the vector is: $5 \times N_b = 320$.

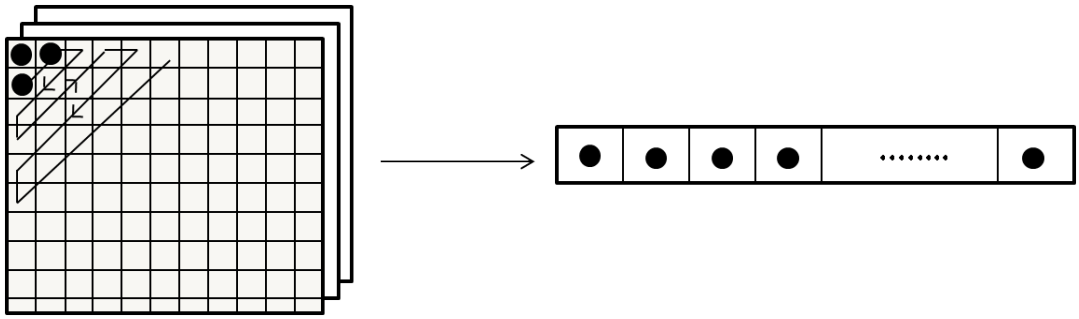


FIGURE 3.6 DCT coefficient extraction in zig-zag pattern from each block. Each image is resized to $64x64$ and divided in blocks of size 8×8 for DCT features extraction.

Experiments show that in case of misaligned images DCT coefficients do not perform up to the mark since there is a lot of frequency variation between consecutive images. However, advantage of using DCT over PCA is two fold (1) size of the feature vector is independent of the number of persons in the database and (2) subspace learning is not required in this case.

3.3.3 Local Binary Pattern

Local Binary Pattern (LBP) [AMH+06b] is an image coding representation for object classification. LBP is introduced by Ojala et. al. [OPH96] as a texture descriptor. A general LBP evaluation criteria is represented by notation (P, R), where P denotes neighborhood and R denotes radius. We calculate features with $(8, 1)$ configuration. Since it is applied on a gray image, we obtain a feature vector of 255 length. Figure 3.7 shows this process in detail.

For a given labeled image $f_l(x, y)$, histogram based feature vector is obtained using

$$H_i = \Sigma I f_l(x, y) = i, i = 0, ..., n - 1 \tag{3.6}$$

Where n is number of different labels. We use $n = 256$ for $(8, 1)$ configuration throughout our experiments.

3.3.4 Sparse Local Descriptors

Current research shows that local representation of the images is more effective as compared to global representation of the images. Image sparseness can deal with several challenges like occlusions, efficiency and requires less memory to process. We experiment this property

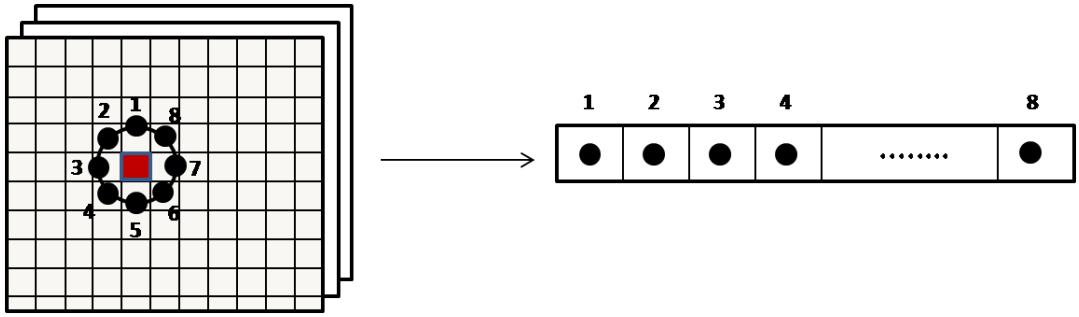


FIGURE 3.7 Local binary pattern with eight neighborhood

by using grid-based points and interest points. The proposed approach is useful because: 1) instead of using full image, it is useful to use sparse representation. This reduces the size of feature vector while preserves the strength of their representation for the given objects. 2) Features are efficient to calculate and have capability to deal with occlusions [WMY+08].

However for grid-based features, the flexibility of grid size and rectangle descriptor size is a trade off between efficiency and accuracy. For a given image I we define a grid of fixed size. For instance, on AT and T database where each image is 112×92, we use a grid 28×23, whereas on Yale-B database where each image is 192×168, we use a grid of 24×21. On each grid point gradient is calculated both in horizontal and vertical direction. In order to get optimal representation, we control the orientation of the rectangular descriptor by using gradient values at these interest points. If gradient value at an interest point in horizontal direction is greater than that of vertical direction, longer side of the rectangle reside along x-axis. These rectangular features are concatenated in a single vector which is the representative of the given image I. We use PCA for feature extraction. The results are shown in section 3.9. Figure 3.8 shows the detailed process of the feature extraction. Similarly for interest points other than grid points, we use same approach. The comparison of the performance on AT and T database database for grid-based and interest points based recognition is shown in Figure 3.10.

3.4 Features Classification

In order to prove the strength of the representative features, we use different classifiers. We mainly, experimented with *Decision Trees* (DT) and *Bayesian Network* (BN). DT uses J48 as the baseline algorithm which implements C4.5. The detail about tree based classifier is given in Table 3.1. For *Bayesian* network we use specification given in Table 3.2.

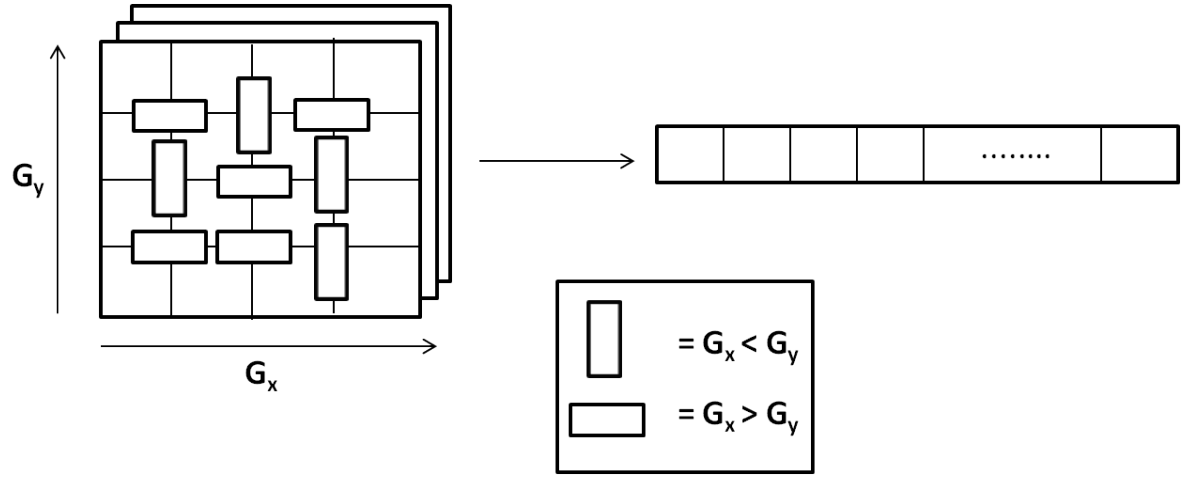


FIGURE 3.8 Gradient based feature extraction. On fixed location gradient values are calculated in horizontal and vertical directions. Gray values from these points are extracted in a rectangular block with larger side in the orientation of larger gradient. Different sizes of rectangular blocks are extracted.

Attribute	Value Assigned
Confidence Factor	0.25
Min- number of instances per leaf	2
number of folds	3

TABLE 3.1 Specification of decision tree used in classification [WF05].

3.5 Graphical User Interface

We develop a *Graphical User Interface* (GUI) for demonstrations of the aforementioned techniques, quick experimentations and providing a tutorial for understanding face image analysis tools. Figure 3.9 shows a simplified view of the GUI used in our experiments. It provides the option to capture images either from videos or from camera. Further, captured images can be processed for local feature detection in real time. A hidden image normalization module rectify the input image and stores features and corresponding label for training the PCA based subspace. Finally, face recognition module classify the images and prints the name of the person on the image to display.

Attribute	Value Assigned
Estimator	SimpleEstimator
search Algorithm	K2
ADTree usage	false

TABLE 3.2 Specification of Bayesian network used in classification [WF05].

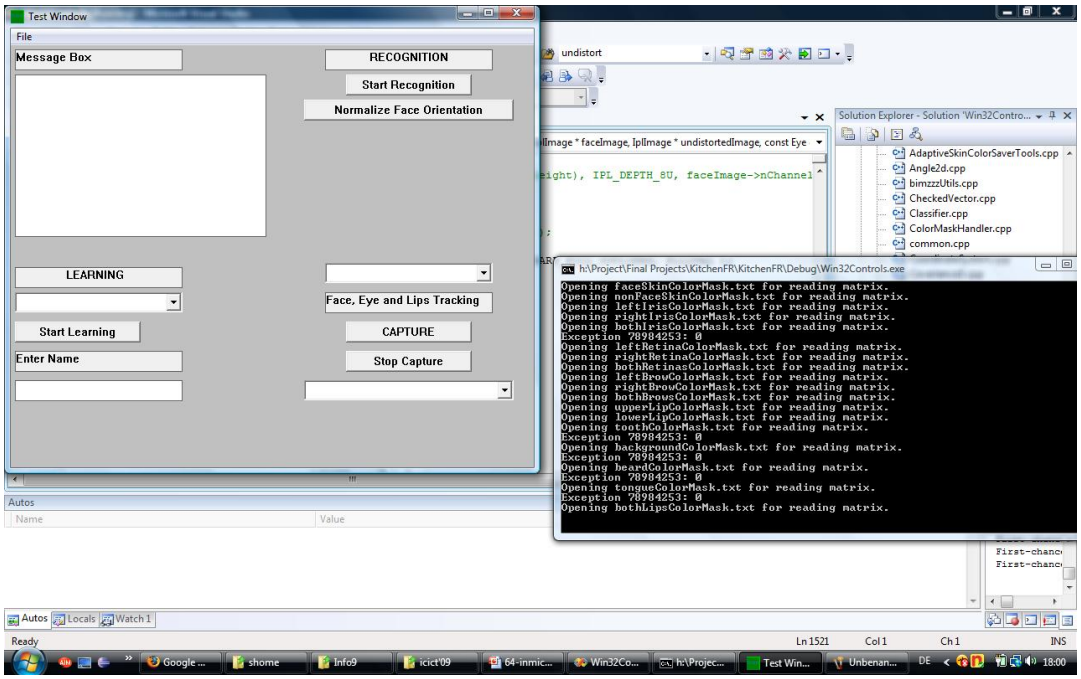


FIGURE 3.9 GUI for face recognition system.

3.6 Interest Points Detection

In order to investigate in detail, we study interest point detection which is one of the widely studied approaches in computer vision from last few years. This approach highlights the points in a given image which are important for further analysis. These points might refer to shape, color, texture, corner or edges in a given image. Different researches have developed approaches for interest point detection and have given different names, for instance, key point detection, salient points, interest points etc. [Tey08]. Such approaches lie in the area of *content based image retrieval* (CBIR) where it is important to find the information inside an image without any prior knowledge about the content of the image. Local feature descriptors extracted from the interest points are more useful as compared to global representation due to:

- Sparse Representation: It reduces the size of the feature vector which not only enhances classification but also provides sufficient robustness.
- Detection and Tracking: It requires no object detection and tracking rather represents only points of interest.
- Object Segmentation: In order to avoid background clutter, objects are usually segmented. However before applying interest point detection, it is not necessarily required to segment the object.
- Occlusions: Sparse representation have proved its strength to deal with the occluded objects.
- Generalization: Interest point detection is a general solution and applied to any image. They are, in general not designed to apply on specific objects.
- Cluttered Images: These points have the ability to cluster around the objects present in an image hence have ability to deal with cluttered background.

Besides several advantages, there are a few challenges in local feature description approaches.

- Loss of Information: Interest points might not represent few parts of an object which might be useful for object classification or provide any other useful information about the object.
- Moving Backgrounds: Spatiotemporal interest points are not stable against moving backgrounds and decrease the classification rate.
- Spatial Information Loss: Generally, interest points lose spatial information of the feature descriptor. However, recent research works have provided several approaches to consider spatial information.

We devise a general framework to find interest points from a given image by assuming that there is at least one major object present in the image. These soft-constraints can easily be extended to more objects. In order to study the representation of these interest points, we consider face recognition as a case study. These points are extracted from entropy coded images. Figure 3.10 shows the comparison on choice of number of interest points. We use regular grid and entropy coded interest points on AT and T database for face recognition.

Local features are extracted from region around the interest points and PCA is applied for final feature set extraction. For details, refer to section 3.9.

Generally, a texture descriptor has three properties: 1) robust to environmental changes (like illumination and geometric variations), 2) rich informative description and 3) efficient in computation. A sparse texture descriptor in our case fulfills these characteristics. The basic aim of this approach is to develop a texture descriptor which is capable to perform in content based image analysis problems.

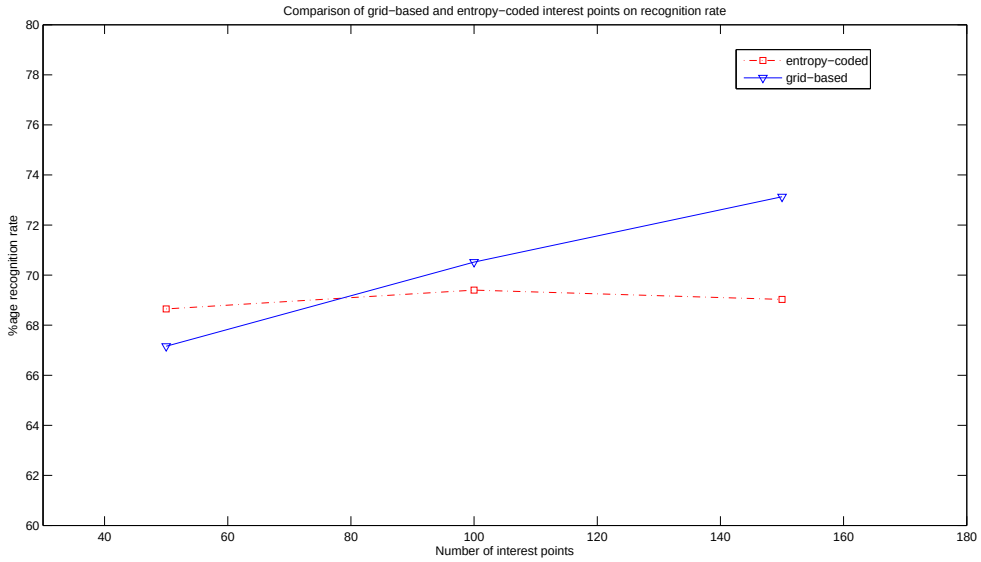


FIGURE 3.10 Effect of interest points on the recognition rate. This explains that our interest point detector finds more meaningful points which are better in recognition. Results are obtained from ORL database.

3.7 Interest Points from Entropy Images

A measure for the uncertainty associated with a random variable is the *Shannon Entropy* or *Information Entropy*. The Shannon Entropy of a discrete random variable X, that can take on possible values is $X_1, \ldots, X_n$ defined as

$$H(X) = E[J(X)] \tag{3.7}$$

i.e. the expected value E of the information content $J(X)$. The information content of a random variable X is associated with the probability of the occurrence of this event and thus

modeled as

$$J(X) = log\left(\frac{1}{p(X)}\right) = -log(p(X)) \tag{3.8}$$

It is typically measured in bits, i.e. the logarithm to the base of two is taken. Substituting 3.8 in 3.7

$$H(X) = -\sum_{i=1}^{N} p(X_i)log(p(X_i)) \tag{3.9}$$

Entropy coded images are obtained by calculating entropy values in horizontal and vertical direction at each pixel value of the image in a small rectangular patch. A larger value of the entropy in a rectangular block is assigned as a value to this pixel. In this way image is coded with maximum disorder at local levels. Higher entropy values indicate less uniformity and helps to filter out uniform areas in an image. Further this entropy coded image is thresholded to obtain higher entropy values. We consider only 15-20% of the highest values in this image. Figure 3.11 shows entropy coded image and thresholded images. Table 3.3 shows examples of some objects other than faces with interest points using this approach. Further examples are shown in Appendix-A.

Since at this stage objects with maximum disorder in the image are prominent hence it is useful to apply a corner detector which further highlights high gradients. These corners mostly reside along the objects present in the image. We cluster all these corners by using k-means cluster. Since it is assumed that at least one major object is present in the image, so we assume that $k = 1$. Finally the corners in the vicinity of the centroid of the cluster are taken as interest points. Detailed flow of our approach is shown in Figure 3.11.

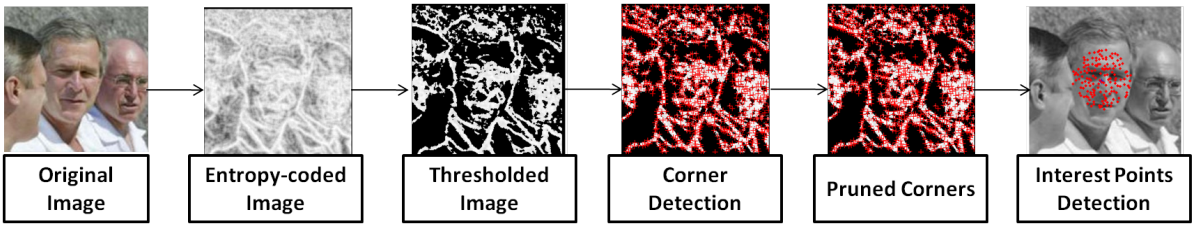
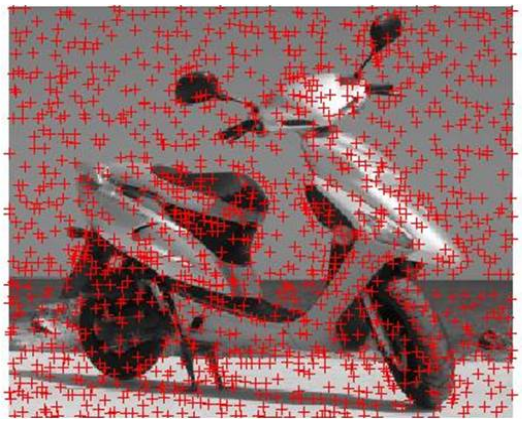
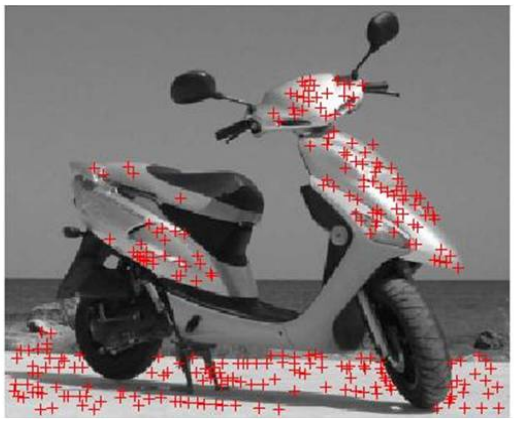
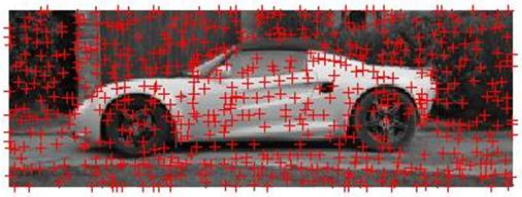
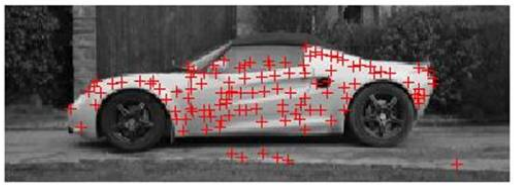


FIGURE 3.11 A sequence of interest point detection from entropy coded images.

a)

b)

c)

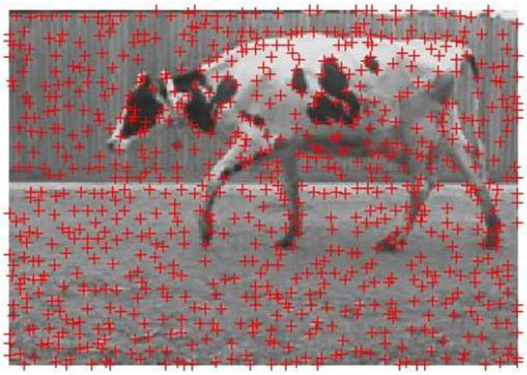



TABLE 3.3 (Left column) *Harris* corner detector applied to original images, (Right Column) *Harris* corner detector applied to entropy coded images.

3.8 Facial Expressions Synthesis

In addition to face recognition, we perform slight facial expressions which are synthesized on a humanoid robot, iCub [Rob]. The system works with eyebrow motion control. This is performed using locations of the local facial features and applying some constraints on their motion. Figure 3.12 shows the functionality of this system. Eyebrows motion is observed by the robot in consecutive image frames and if the motion is recognized as a valid expression by the robot then robot starts following it from interacting persons. In case both eyebrows move up relative to the a reference point, it is recorded as a *surprise* expression. The system can also classify individual motion of the eyebrows which can be seen in Figure 3.12.

FIGURE 3.12 Facial expressions controlled by eyebrow motions.

3.9 Experimental Evaluations

In this section, we provide two different types of experiments. In section 3.9.1, we study experimentation performed during real time face recognition applications. In section 3.9.2, we describe experiments related to some of the novel approaches used in this thesis.

3.9.1 Experiments for Face Recognition

In real time face recognition system, we capture multiple images in a sequence. Although no temporal component of the facial variation is recorded for recognition, but most of the training and testing samples are continuous in time. We discard those frames where facial features are falsely registered. Feature localization is fully automatic and restricted with some constraints to avoid registration errors. We capture 60 images and register them to a standard template. For every input image, further images are synthesized by choosing four additional points in

the vicinity of the detected position of the lips. In this process we get more images than the training examples. We use k-means clustering to get same number of images back by keeping $k = 60$. PCA and DCT features are collected from these images.

All the images in the training set are normalized (using procedure in section 3.2). Three different types of features are calculated from these normalized images. These features are mentioned in detail in section 3.3. The system can perform on a limited set of faces with hundred percent accuracy. We experimented it in different demos on four to five different persons successfully [Ria08][Ria09][Ria10]. From different feature sets, we perform recognition with different distance classifiers working in parallel. If the results of all classifiers match, the given identity is displayed with the image. Further evaluation of this approach is given in next section on Yale database. In next section, we also observe that sparse features perform better than holistic approaches.

3.9.2 Experiment for GB features

In order to test the performance of our proposed approach of gradient based blocks, we use two different databases for face recognition. The choice of these databases corresponds to their similarity to the image normalization process explained in section 3.2. These databases are *AT & T* and *Yale B*. Former consists of images from 40 persons with ten images per person in the database. The variations in the images consists of slight out of plane rotation and glasses/without glasses, lighting variations, slight facial expressions and open/closed eyes. The size of each image is 112×92 stored as gray scale image. On the other hand Yale B database is relatively larger database with 10 persons in the database where each person has 65 images. These images correspond to 9 different viewing conditions and with 64 different illuminations with one ambient image. The database contains in total 5760 images.

For experimentation, each database is randomly divided in two parts. One third of the images are used for subspace learning and remaining two-third of the data is used for testing. Block based gradient oriented features are calculated with a fixed grid and fixed window size (as explained in section 3.3.4). PCA is applied on these sparse raw feature vectors as explained in section 3.3.1. For testing purpose, the remaining two-third data is further divided in two parts. Two-third of this remaining data is used for training the classifier for face recognition and reaming part is used in classification. Since each database is randomly divided in two parts, so there may be a variation in the results due to different choices of images. In order to deal with this issue we repeat our experiment five times and averages the results. However each attempt shows almost comparable results. The size of the feature vector for Yale database is 685 while for AT and T database is 472. The size of the feature vector shows the

Approach	DCT Coefficients	Holistic Eigenfaces	Grid Based Sparse Features
Recognition Rate	72.04%	94.48%	98.04%

TABLE 3.4 Comparison on results on Yale database for sparse gradient based features and conventional eigenface approach. The results show that instead of choosing the whole image for feature extraction, it is recommended to use sparse features extracted from gradients with their spatial relation. The overall recognition rate is extracted by averaging the classification rate after conducting the experiments five times for uniformity in results [LHK05].

trade off between representation power and compactness. For gradient based approach, we get less compact feature vector but representing higher recognition rate however in case of traditional eigenface approach we get compact feature vector but with relatively less recognition rate. In both cases we preserve more than 99% energy of covariance matrix.

Since feature set arise only from one source hence Bayesian network is used as a classifier for the experiments. Other classifiers can also be used in this regards. The results are shown in Table 3.4.

3.9.3 Interest Points and Window Size

Although feature set presented in section 3.3.4 are efficient and provide strong representation of a single image however the choice of grid size and rectangular window size is kind of an art. Grid size is chosen by trading off between the size of the image and the size of the resulting feature set. A grid is chosen so that the rectangular windows should not overlap with each other and resulting feature vector size does not increase from the original image. For the rectangular window, a block of different width and height is chosen. In order to verify this fact, we used a square instead of rectangle with sides equal to the greater side of the rectangle. The performance of the system degrades. The fact behind is the loss of strength of the features caused by gradient based orientation. The results are shown in Table 3.4. This emphasizes that the size of the block should be rectangular and oriented by the gradient direction in 2D image space.

3.10 Summary and Conclusions

This chapter provides methods for the development of a real time face recognition system. It also servers as a self-contained tutorial for an unconstrained face recognition system. We have studied face detection, facial features localization, image registration, feature extraction and finally classification. For a given image, a face detector followed by a set of local facial

feature detectors is applied for image normalization. Face detection module uses standard Viola and Jones [VJ04] detector whereas eyes are searched in a detected face window using template matching and lips are detected using color information. These three points are used to register a face image to a standard template. Image registration process helps to normalize the images against variations in different view points and improves the classification rate. We study three different feature sets, which are 1) *holistic features* using PCA, 2) *local descriptors* which utilize DCT coefficients and LBP coding and finally 3) *sparse features* using grid points and interest points from entropy coded images. The results on AT and T database and Yale-B database show that the sparse features perform better as compared to the other two approaches. These facts have also been studied by the other researchers under occlusions [WMY+08]. We study sparse features using two different approaches, which include 1) grid based features and 2) interest point detectors. Both approaches outperforms DCT, LBP and PCA for face recognition on Yale-B database . Furthermore, we also study facial expressions synthesis on a humanoid robot using eyebrow positions. Finally we conclude this chapter with the following observations:

- Conventional image based approaches for face recognition are efficient but exhibit limited classification results against current outstanding challenges like varying facial poses, expressions and illuminations. A thorough pre-processing and normalization of the face images can improve the classification rate by making the system capable for real world applications.
- Different types of detectors work in parallel to detect faces, eye positions, eyebrows, nose and lips which shows their potential to work in real time for image registration.
- A comparative study of different feature set for face recognition has been conducted with intensive experimentation, which shows that sparse features have potential to perform better recognition.
- We study interest points extracted from entropy coded images. These interest points represent high entropy values at different locations and perform better as compared to original images. This can be seen in Table 3.3.
- Facial expressions are synthesized from the simple distance constraints on the local features, which works successfully on a humanoid robot (see Figure 3.12).

Such face recognition systems can work efficiently and robust to slight variations in the face images. However large rotations in head pose may lead to false results in finding the

facial features, which reduces accuracy. From this chapter we can conclude that the content based approaches have shown their potential for unconstrained face recognition applications. Furthermore, sparse approaches outperform conventional image based approaches. As a future work of this research, we suggest to use feature descriptors extracted from interest points of the normalized images. This improves the recognition rate and also performs satisfactorily in real time. In next chapter, we study 2D face modeling which provides better image registration by using more points and improved recognition rate. Model based approaches are relatively slower than image based approaches studied in this chapter, however are more reliable and robust for real world applications.

CHAPTER 4

2D Face Modeling

This chapter addresses modeling of the human faces in 2D space. We mainly focus on *Active Shape Models* (ASMs) and *Active Appearance Models* (AAMs) . Our goal is to extract a set of useful and robust features for different face classification applications. The feature set is extracted from the face image sequences and efficiently utilized for face recognition, facial expressions recognition, gender classification and age estimation (for details refer to section 4.7). The approach follows in: (1) calculating AAM parameters from the face images, (2) using optical flow to calculate motion features for facial expression, (3) and finally using this feature set for classification purposes. The major contribution of this chapter is to configure a spatiotemporal feature set which comprises of shape, texture and temporal deformation parameters of the face in a given image. Furthermore, this feature set is experimented for face recognition in the presence of different facial expressions. Our approach is image based and has also been applied to videos and image sequences. Since the model fitting does not require detailed texture information hence the proposed system is efficient to work in real time.

Depending upon the nature of the feature components, we can divide them as geometrical features, textural features and temporal features. A vast literature of face image analysis over the last few years (details in chapter 2) shows that face recognition systems require essential texture and structural components [CET01][BV03b][LTC95b][ETC98]. Facial appearance with adequate local appearance plays a significant role in recognition, verification and watch list check scenarios. On the other hand, facial expressions are usually person-independent and a facial expressions classification system requires shape and temporal variations [MWS+08][WRMR08][WMSR07]. An additional facial appearance results an improvement in expression classification at the cost of more computations. Similar to face recognition system, gender classification requires geometrical and textural variations to classify between two gender classes [RMBR09c][RGBR09]. On the basis of the aforementioned knowledge and experimental research of over a couple of decades, we can configure a feature vector which

can represent multiple facial attributes in a compact form. We term this feature vector as *Spatiotemporal Multiple Feature* (STMF) . Table 1.1 in Chapter 1 summarizes the significance of these constituents of the feature vector with their primary and secondary contributions toward the feature set formation. Since our feature set consists of all three kinds of information, it can represent spatial and dynamical patterns in a compact set of parameters. The proposed feature vector is then successfully experimented for facial identity, facial expressions recognition, gender classification and age estimation. Model parameters are obtained in an optimal way to maximize information within the face region in the presence of different facial variations. The robustness of the feature vectors against real world challenges has also been studied during the experimentation.

4.1 Background

The future of the interactive technologies relies on the real time interpretation of the human body language and its analysis for perception learning. In such scenarios, machines might play an essential role in the individual's life and one might often be confronted with situations where humans and machines are interacting with each other and performing joint activities. This requires to build the intelligent and user-friendly systems to ensure a seamless interaction between a human and a machine. One of the good examples of such systems is an assistive robot [ea07] which is capable to serve like an attendee nurse to elderlies, in order to safely perform their daily life activities. These intelligent machines, also called robots should be able to interact with humans of all categories (e.g. ages, gender and ethnicities) and besides training persons these systems should perform equally good for untrained and new user. Machine intelligence can be measured from its perception about the environment and manipulation capabilities without human intervention.

For this purpose, an intuitive approach is to borrow the concepts of human-human interaction and train the machines for comparable performance in real world situations. Human faces play an important role in daily life interaction. Therefore, context and identity awareness capabilities of a robot improve their performance under joint activities. A system aware of person identity information can better utilize user specific habits and can store person dependent knowledge for improving future interactions.

In the remaining part of this chapter, we study a brief introduction about the modeling object in computer vision; their representation and parameterization. In section 4.2 three different model fitting approaches has been studied. We study conventional ASM fitting, objective functions based approach for 2D face models and finally inverse composition image alignment

approach, which is one of the most efficient approaches in the literature of model fitting. We further explain basic 2D geometric and textural models including active contours, ASM and AAM in section 4.3. In section 4.4 a method for temporal feature extraction is studied using optical flow. In section 4.5 a detailed description about model based image interpretation is given. We explain the contribution of different features toward the formation of a single feature vector for classification. In section 4.7 we thoroughly provide the experimentation performed using appearance model. Finally section 4.8 concludes the chapter with results and implementation scenarios of this work for real world systems with recommended extensions of this work.

4.1.1 Modeling Objects in Computer Vision

In computer vision, a model is defined as a compact representation of an object. A model may be a geometric model, color model, motion model or a probabilistic model. In general, models are context-aware and object specific however they can be extended as generalized models to a given class.

Human face modeling has been one of the widely studied topics in the area of face tracking, HRI applications, biometrics, behavior analysis and facial animation. By the development of efficient and fast algorithms, availability of better hardware and their capability to deal with real world scenarios, currently face modeling is one of the challenging fields in computer vision. After the failure of Bertillon system [Ber09] in face recognition system, researchers started paying attention to develop a reliable system to recognize the humans from their faces. This was followed by a research and effort of several years which finally resulted in different commercially available face recognition systems. Face modeling started roughly in mid eighties [WH04]. A face model in general comprises of the structure of the face. This structure can either be defined using contours [KWT88][GRCC04] or anatomical landmarks [CET98a]. Note that these anatomical landmarks and contours are interrelated. This structure defines the shape representation of the local facial features, which can vary from person to person and under varying factors. Another important part of the models is the color information. Color information can be defined as simple gray values [LTC95a], face texture [RMW+09] or detailed texture map in computer graphics and games applications [Bus][RGBR10]. The model of an object is generally a small set of parameters to control the variations of this particular object in a relatively low dimensional subspace. The set of parameters is constrained to the degrees of freedoms of the object. This filters out unrealistic and hallucinated views of the object. In other words, the selected parameters correspond to the realistic and meaningful motions of the object. A simple model can be represented as:

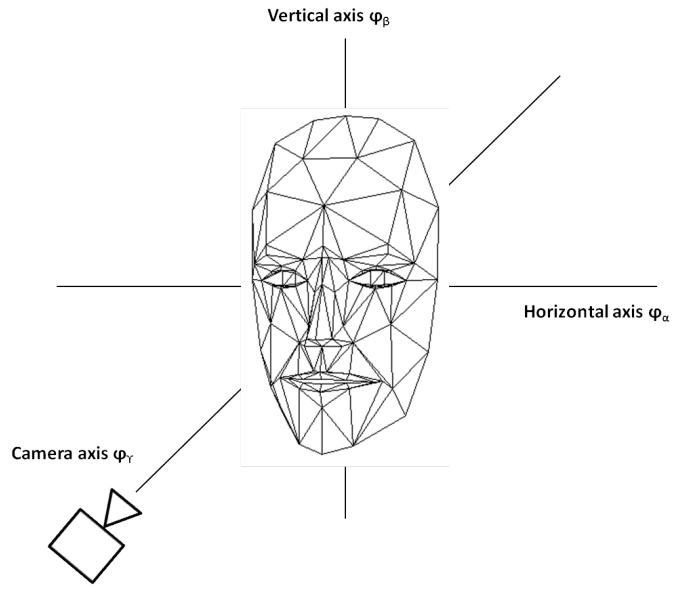


FIGURE 4.1 An example of a 3D face model. The model can perform six global motions include translation and rotation in three axis which are governed through model parameters.

$$p = (s, r, t, x, g) \tag{4.1}$$

Where components of the models are the parameters showing different variations. For instance, s is a vector of parameters controlling the scale, r and t are rotation and translation vectors respectively, x and g are shape and textural parameters respectively. These parameters can be time dependent or independent. A parameter set changing with time is more complex and defines the deformation of the object. The parameters that perfectly define the object in the image are the optimal parameters p^* and referred to as *ground truth* for that image. In 3D space, an extra dimension is added to the parameters. The translation vector can translate the model in three dimensions $t = (t_x, t_y, t_z)$. Another important motion is the 3D rotation which is defined with $\Phi = (\phi_\alpha, \phi_\beta, \phi_\gamma)$ where ϕ_α, ϕ_β and ϕ_γ are rotation around horizontal axis, vertical axis and around camera axis respectively.

On the basis of their characteristics, the models are divided in generative models or discriminative models. A generative model represents a set of hidden variable which describes different semantics of the objects. These hidden variables are capable to reproduce observed data at any time. AAM, for instance are generative models. However, if a model represents some hidden variable but the reproduction of the observed data is not possible then the model is called discriminative. Neural networks, for instance are the example of discriminative models.

Further, on the basis of their functionality, models can mainly be divided in two types, rigid and non-rigid model or deformable models. Rigid models define a compact shape and texture of the object which cannot be varied due to local structural and textural variations of the object, examples include [May07][WMSR07]. Rigid models can be synthesized to global variation of the objects. On the other hand, deformable models are the models which can be adapted to different local and global variations of the object, examples include [CTCG95][CET98b][RMW+09] [Ahl01][BV03b][WH04].

The structure of the model can be defined by using three different approaches as suggested by [CTCG95]. All these approaches define the landmark points with different intuitions. These include: (1) *Anatomical Landmarks:* These landmarks represent the points on various locations on an anatomical objects like face, hand or full body. (2) *Mathematical Landmarks:* These landmarks correspond to some points which are of mathematical importance for example edges, corners or blobs. (3) *Pseudo Landmarks:* These landmarks are constructed between anatomical or mathematical landmarks to get more sample points for analysis. They are defined with some rule or generally equally spaced.

4.1.2 Model Parameterization

In common practice, model parameterization is generally performed using PCA. It is adapted by ASM, AAM, morphable models and Candide model. If X_i, $i = 1, \ldots, N$ is any vector, where N is the total number of vectors, we parameterize it by using mean and covariance matrix. This vector X_i may represent structural data, textural data (gray values or RGB color values) or temporal velocity vectors. The mean vector $\overline{X}$ is calculated as:

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{4.2}$$

and the covariance matrix C is given by:

$$C = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x}) (x_i - \overline{x})^T \tag{4.3}$$

Eigenvectors ϕ_t and their corresponding eigenvalues λ_t of the covariance matrix C are calculated. These eigenvectors are sorted in a descending order corresponding to their eigenvalues. Only n significant eigenvalues are considered. Choice of n depends on designer behalf or in general it is taken as top $5 - 10\%$ of the highest eigenvalues. If P is the matrix of eigenvectors such that $P = [\phi_1, \ldots, \phi_n]^T$ then we can write:

$$X \approx \overline{X} + P * b \tag{4.4}$$

Where b is an n-dimensional parameter vector and can be calculated by using the fact $PP^T = I$ (since eigenvectors form an orthonormal basis).

$$b \approx P^T \left(X - \overline{X}\right) \tag{4.5}$$

We use equation 4.5 to find parameters of shape, texture and temporal features. General methodology used during our experiments is to divide the data in two different parts. One-third of the data is used for subspace learning to form an eigenspace. The remaining two-third of the data is then projected to this space and the weight matrix is calculated in eigenspace. These weights act as the feature vector. In addition to PCA parameters, we have also extracted texture parameters using DCT, LBP and BoW. DCT is applied to texture map in a zig-zag pattern to each block. From all 184 block (where each texture block represent a triangular patch on face surface), five DCT coefficients are extracted from each block making a feature vector of size $184 \times 5 = 920$. The choice on number of DCT coefficient is taken from common practices in literature and trading off between accuracy and efficiency [ESTS09]. PCA is applied to reduced the dimensions of this vector. For LBP, an image is coded at each pixel inside face region. A standard LBP coding with $(8, 1)$ configuration is used to generate a feature vector of length 256 codes.

Since shape and texture parameters arise from two different sources, it is not quite obvious to combine them together. There are different approaches to do this task. A simple way is to normalize and concatenate the feature vector together. In this way, the contribution of each feature is weighted equally and observations with high values will not dominate those with the lower values. Cootes et al [LJ05] use weight matrix to configure combined AAM. Since b_s and b_g are shape and textural parameters respectively, hence

$$b = \begin{pmatrix} W_s b_s \\ b_g \end{pmatrix} = \begin{pmatrix} W_s P_s^T (x - \overline{x}) \\ P_g^T (g - \overline{g}) \end{pmatrix} \tag{4.6}$$

where W_s is a diagonal matrix of weights for each shape parameter, compensating for the difference in units between the shape and gray models. A further PCA is applied to remove any correlation between these two modalities.

$$b = P_c C \tag{4.7}$$

Approach	Introduced by	Pros	Cons
ASM Fitting	Cootes et al, 1998 [CTCG95]	Simple to calculate	Not robust, slower, converge to local minima
Objective Function	Wimmer et al [WSPR08]	Model Fit to novel images, robust, automatic learning	Faster than ASM but slower than ICIA algorithm
ICIA fitting	Matthews et al [MB03]	Fast, robust	Fitting is only possible to those poses which are available in training set.

TABLE 4.1 Comparison of different model fitting methodologies.

Where P_c is the combined eigenvector and C is the parameter vector of texture controlling both shape and texture. For further detail, refer to [LJ05].

4.2 Model Fitting

Fitting a model to a face image is to find the best set of parameters which describe the shape of the given face. This best parameters set is also called ground truth. In other words, goal of the model fitting is to find the optimal parameters which are as close as possible to the ground truth. The fitting problem becomes more challenging in real world application where ground truth value is not available and required to be approximated using the context information. We study three different model fitting methodologies which are efficient and widely used for fitting.

4.2.1 ASM Fitting

ASM is a generic model and is fitted to any new shape using least square solution proposed by [CTCG95]. If S is the simple similarity transformation then a given shape X can be fitted to any shape Y by reducing the least square error e_{pts}:

$$e_{pts} = |Y - S(\overline{x} + \Phi_i b_i)|^2 \tag{4.8}$$

Where Φ_i and b_i are i^{th} eigenvector and shape parameter respectively. More generally, weights can be individually assigned to different points. If W_{pts} is the weight matrix then we can write:

$$e_{pts} = [Y - S(\overline{x} + \Phi_i b_i)]W_{pts}[Y - S(\overline{x} + \Phi_i b_i)] \tag{4.9}$$

An approximation is applied by considering that the global transformation repetitively oc-

curs in two steps until convergence: (1) Solve for the pose parameters t assuming a fixed shape b_s, (2) Solve for the shape parameters b_s, assuming a fixed pose.

For further details refer to [LJ05]. A rather more efficient approach was introduced by Baker and Matthews by using inverse compositional image alignment which is explained in detail in section 4.2.3.

4.2.2 Objective Functions Fitting

We use a method developed by Wimmer et al [WFS+08] for face model fitting. This methodology uses the design of an objective function for model fitting. An objective function may correspond to any mathematical function like energy function, cost function, likelihood function and quality function. In this approach a minimum value of the objective function describes the best fitting results and hence objective function corresponds to cost function. Objective function value at each fiducial point ranges between $[0, 1]$. A fitting algorithm searches for the optimal parameters which minimizes the value of the objective function. For a given image I, if $E(I, c_i(\mathbf{p}))$ represents the magnitude of the edge at point $c_i(\mathbf{p})$, where $\mathbf{p}$ represents set of parameters describing the model, then the objective function is given by:

$$f^{\theta}(I, \mathbf{p}) = \frac{1}{n}\sum_{i=1}^{n} f_i^{\theta}(I, c_i(\mathbf{p})) = \frac{1}{n}\sum_{i=1}^{n}(1 - E(I, c_i(\mathbf{p}))) \tag{4.10}$$

Where $n = 1, \ldots, 134$ is the number of vertices c_i describing the face model. An objective function design is proceeded by annotating few training images with preferred model parameters. This step is the only manual intervention and the remaining steps are automatic. An objective function is subjected to three constraints:

- R1: The objective function $f(I, \mathbf{p})$ has a global minimum at $\mathbf{p} = \mathbf{p}^*$.
- R2: The objective function $f(I, \mathbf{p})$ has no local extrema or saddle point at $\mathbf{p} \neq \mathbf{p}^*$.
- R3: At any $\mathbf{p} \neq \mathbf{p}^*$ the slope falls most toward $\mathbf{p}^*$. Therefore, the direction of the gradient $\nabla f(I, \mathbf{p})$ needs to point away from $\mathbf{p}^*$.

During the training phase f^{θ} is split up in local parts f_i^{θ} at each point. This splitting serves as an approximation to the main objective function.

$$f^{\theta}(I, \mathbf{p}) := |\mathbf{p}^* - \mathbf{p}| \approx \frac{1}{n}\sum_{i=1}^{n} |c_i(\mathbf{p}^*) - c_i(\mathbf{p})| =: \frac{1}{n}\sum_{i=1}^{n} f_i^{\theta}(I, c_i(\mathbf{p})) \tag{4.11}$$

the overall design of the objective function is divided in four steps:

- Image annotations: This is the only manual step and faces are annotated by the user with the preferred model. This preferred model may also be a ground truth for the given image.
- Generating annotations: From these manual annotations, further annotations are automatically generated which reduces the effort in learning the objective function. These automatic annotations are restricted to the line perpendicular to edge at point $c_i\{\mathbf{p}\}$.
- Features specification: Training data is generated by extracting features at different location along the perpendicular line. The feature extracted by Wimmer et al are Haar-like feature from integral images. Haar-like features have benefit in these scenarios since they are stable against noise, efficient and allow to model the border at non-distinctive borders.
- Training: Finally model trees are used to map the extracted features to the values returned by f_i^θ. Objective functions for different contour points use different subsets of Haar-like features.

There are different benefits of using this approach.

- Annotating images is more intuitive and less error-prone than explicitly designing an objective function.
- Model tree selects the relevant features from the large set which are more objective.
- The learned objective function is optimal.

Figure 4.2 represent some of the model fitting results using objective function approach.

4.2.3 Inverse Compositional Image Alignment

Inverse compositional image alignment (ICIA) algorithm is faster than conventional ASM fitting. In this approach incremental warp is calculated using bases appearance. It was proposed by Matthews et al [MB03]. Let $\mathbf{p}$ represents AAM parameters corresponding to the current image I with warp $W(X;\mathbf{p})$. Where $X = (x, y)^T$ defines the base shape S_o. Any change in parameters $\mathbf{p}$ is represented by $\Delta\mathbf{p}$ and the incremental warp is calculated by $W(X;\Delta\mathbf{p})$. On contrary to forward compositional algorithm, the incremental warp is calculated from $A_o(X)$ instead of $W(X;\mathbf{p})$. Where $A_o(X)$ is the base appearance. The minimization is performed over:

FIGURE 4.2 Model fitting results using objective functions.

$$\sum_{X}[I(W(X;\mathbf{p})) - A_o(W(X;\Delta\mathbf{p}))]^2 \tag{4.12}$$

with respect to $\Delta\mathbf{p}$ and then updating the warp using

$$W(X;\mathbf{p}) \leftarrow W(X;\mathbf{p})oW(X;\Delta\mathbf{p})^{-1} \tag{4.13}$$

The proposed approach outperforms the other fitting approaches in the sense of:

- speed of convergence because fewer iterations are needed to converge to any given accuracy
- frequency of convergence, the proposed algorithm converges even from a large distance from the destination. It can be seen in Figure 4.3.
- computational cost because appearance variation is projected out.

Any fitting algorithm can be used for our point distribution models. We implemented later two approaches during our experimentation phase. We prefer to use objective function approach because this approach has the ability to fit to any unknown face, it is real time, objective function is simple and automatic to learn and can work with combined active appearance model.

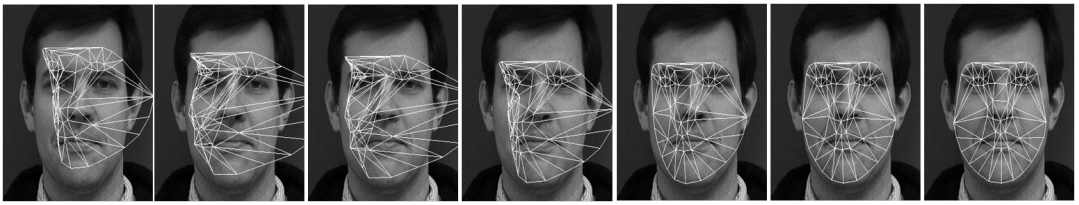


FIGURE 4.3 Model fitting and convergence using ICIA algorithm [MB03].

4.3 2D Object Modeling

4.3.1 Active Contours - Snakes

Active contours or snakes are flexible models which define geometry of the objects in an image. These geometric features include edges and curves with smoothness constraints. Energy spline curves are modeled with stiffness and elasticity. Model parameters are obtained from control points defined on the contours. For details refer to [KWT88]. A snake is an elastic contour which is fitted to features detected in an image. The nature of its elastic energy draws it more or less strongly to certain preferred configurations, representing prior information about shape which is to be balanced with evidence from an image. If inertia is also attributed to a snake it acquires dynamic behavior which can be used to apply prior knowledge of motion, not just of shape. Rather than expecting desirable properties such as continuity and smoothness to emerge from image data, those properties are imposed from the start.

4.3.2 Active Shape Models (ASMs)

Active Shape Models (ASMs) exist with a variety of forms, principally snakes, deformable templates and dynamic contours [BI98]. An ASM consists of average positions of the points obtained from a set of training images and a set of parameters which control different deformation modes. These variations are defined within a given class relative to a given model. These models are context aware and designed by using prior knowledge of the object during the learning phase. An initial guess is assigned in terms of scale, orientation and position in the given image. This guess is then further improved by comparing image data and the initial model fit. Various fitting approaches have been used till date for this purpose. We study three major approaches which are given in Table 4.1. Since ASMs are context aware and use prior knowledge about the object, therefore they exhibit object-specific dynamics. This means that the model dynamics are constrained only to those variations which are present in training images. This is achieved by finding those shape parameters which are uncorrelated. This char-

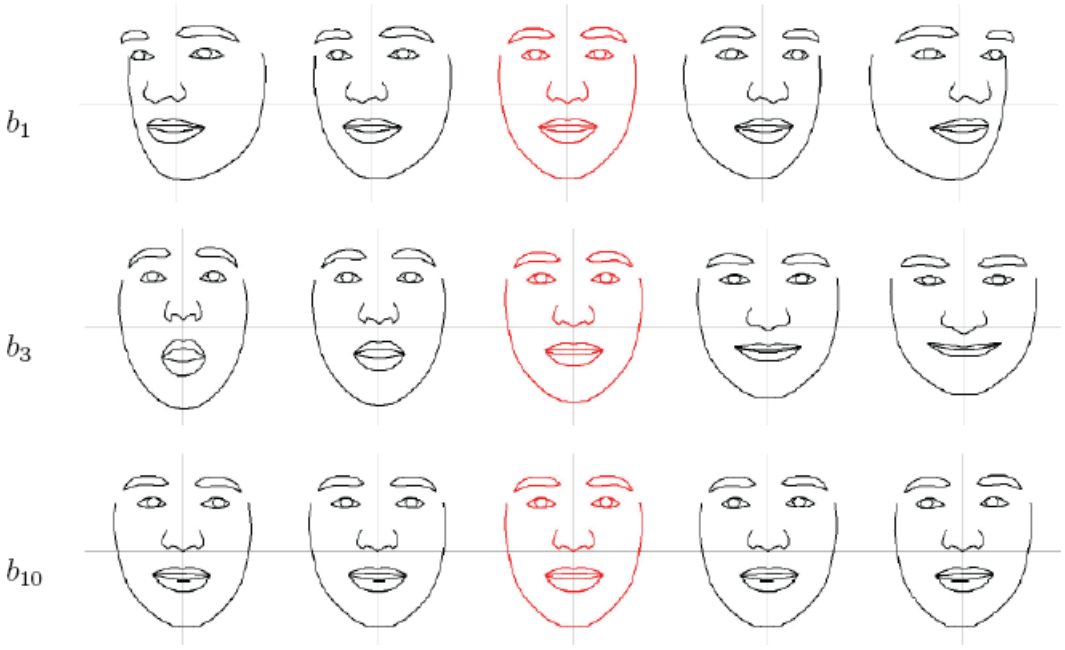


FIGURE 4.4 Facial deformation caused by three different parameters. The deformation is conducted $b_i \pm 3\sigma$, where i is ith parameter and σ is the standard deviation [Wim07].

acteristic benefits in filtering out the non realistic object deformations. For example, human faces can exhibit local and global geometrical variations as a linear combination of the shapes available in the training data. Figure 4.4 shows the motion caused by three different model parameters.

ASMs are inspired from active contours or snakes and were formally introduced by Cootes [CTCG95]. An ASM defines shape of an object in detail. For human faces, an ASM is defined with different number of fiducial points which describe facial features like both eyes, eyebrows, nose, lips and outer face area. These points are optimized with maximum variance using PCA and whole shape is controlled by minimum number of descriptors. These descriptors are parameters of the model which describes various facial deformation. Figure 4.5 shows a generic shape model used in our case with different points and Figure 4.4 shows example facial deformation using three shape parameters. Labeling the fiducial points is not arbitrary rather defined by user or object dependent. For example, in case of human faces, a point distribution represents facial features like eyes, eyes corners, brows, nose and lip corners etc. Further, more number of points add details in defining the model. ASM has been used with 68 points by Cootes et al and Matthews et al [CET98b][MB03]. However, we experimented with the model consisting of 134 points as shown in Figure 4.5. Milborrow et al [MN08] show relation between point-to-point error relative to 68-point-model and number of landmarks. They show a better

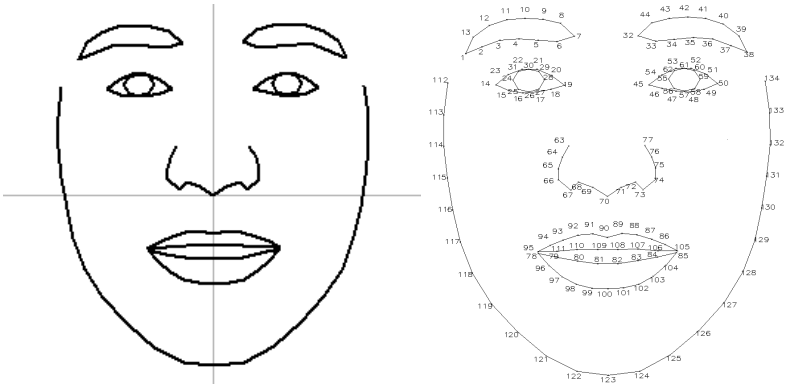


FIGURE 4.5 A generic shape model used in our experiments with 134 fiducial points marked on different facial features [Wim07].

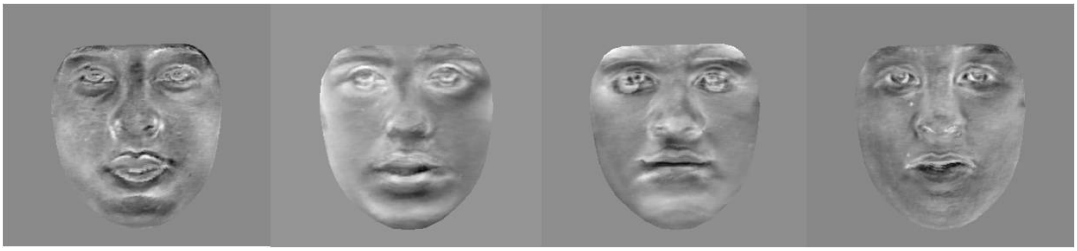


FIGURE 4.6 Different appearance modes of the faces from the database.

performance by using more landmarks in fitting a model.

A point distribution in 2D space forms a convex hull and this distribution can be divided in non overlapping subdivision for mapping the texture between two shapes (one of them is the reference shape which is described in section 4.3.3). Famous subdivision approaches are voronoi diagrams and delaunay triangulations. In delaunay triangulations image plane is divided into a number of non overlapping triangles such that the circumcircle of the triangle contains only the vertices of the triangle and no other point of the subdivision. This is a triangulation which is equivalent to the nerve of the cells in a voronoi diagram, i.e., that triangulation of the convex hull of the points in the diagram in which every circumcircle of a triangle is an empty circle [OBSC00].

ASM fitting is performed by minimizing sum of square distances between corresponding points of the two shapes. This is performed using scaling, rotation and translation mapping of the points.

We obtain 17 parameters for our face model. A shape X is represented in 2D with position of fiducial points e.g. $X = [x_1, \ldots, x_n, y_1, \ldots, y_n]^T$ where $n = 134$ in our case. A shape model is represented with $(t_x, t_y, s, \theta, \mathbf{b})$ where t_x and t_y are horizontal and vertical components of

translation vector, *s* is scaling factor, θ is rotation component and ***b*** is the deformation vector obtained by PCA.

The model is parameterized using PCA to form the shape feature vector.

$$X = X_m + P_s b_s \tag{4.14}$$

Where the shape *X* is parameterized by using mean shape x_m and matrix of eigenvectors P_s to obtain the parameter vector b_s.

4.3.3 Active Appearance Models (AAMs)

An ASM describes structural information however AAM describes combined shape and textural information. Figure 4.8 shows combined shape and texture information for face images. Texture is represented as gray values or color values from the region encompassed by shape model. In order to get a uniform length texture vector from a face region, we use affine mapping which is also adapted by various researchers. The mapping warps texture from a given shape to a reference shape. This reference shape is a standard throughout the experiments. This reference shape can be any shape but in practice it is the mean of all the shapes available in the training data. Texture is extracted from each triangular patch to the corresponding patch of the reference shape. Fiducial points define a convex hull in 2D space and planar subdivision is performed on these points. We perform delaunay triangulation to obtain 236 non-overlapping triangular patches.

Given a set of shape points *X* of the input example image and X_{avg} of the average image, we can find the texture vector g_{im} as follows:

1. Compute the pixel position in the average shape.
2. Find the relative position in example image.
3. Sample the texture values at the points inside convex hull of the average image forming texture vector.

The procedure is shown in detail in Figure 4.7. The texture vector is normalized to remove global lighting effects. This is performed by applying the linear transformation,

$$g = (g_{im} - \beta 1)/\alpha \tag{4.15}$$

where,

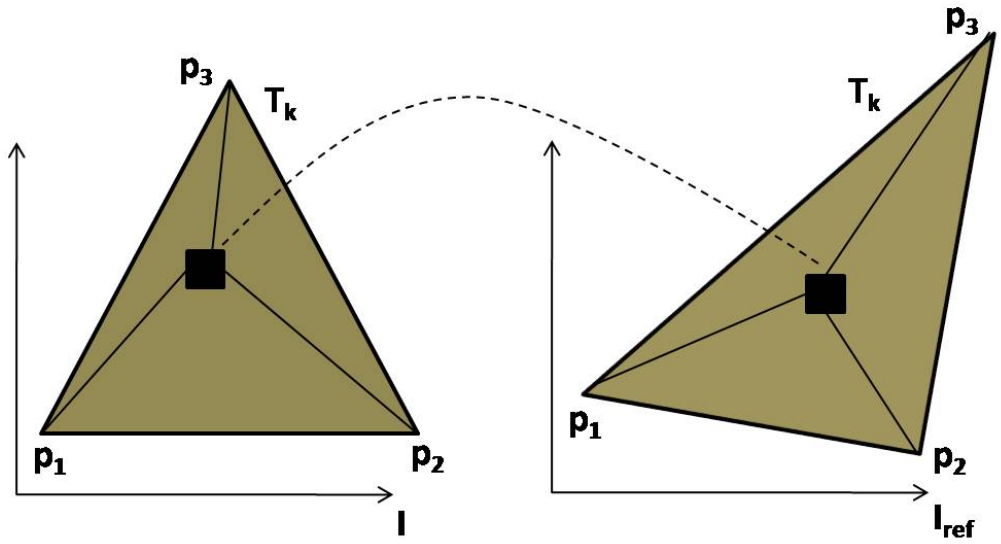


FIGURE 4.7 Piecewise texture warping: For any pixel in a triangle T_k of the reference shape its relative position is calculated in corresponding triangle T_k of the shape of the given face. The pixel value is interpolated and projected to the reference shape. The pixels in each triangle are located by using barycentric coordinates.

$$\beta = (g_{im}.1)/n \tag{4.16}$$

$$\alpha = g_{im}2/(n - \beta) \tag{4.17}$$

Once the texture is extracted it could be parameterized using PCA as,

$$g = g_m + P_g b_g \tag{4.18}$$

Where the texture g is parameterized by using mean texture g_m and the matrix of eigenvectors P_g to obtain the parameter vector b_g. Figure 4.6 shows different facial appearance modes.

Image warping is performed using affine transformation which has the parallelism and lines preservation property. Piecewise affine transform is used to warp the texture of example images on the normalized image. If x_1, x_2 and x_3 are the vertices of a triangle then any point lying inside the triangle can be written as:

$$x = \alpha x_1 + \beta x_2 + \gamma x_3 \tag{4.19}$$

where

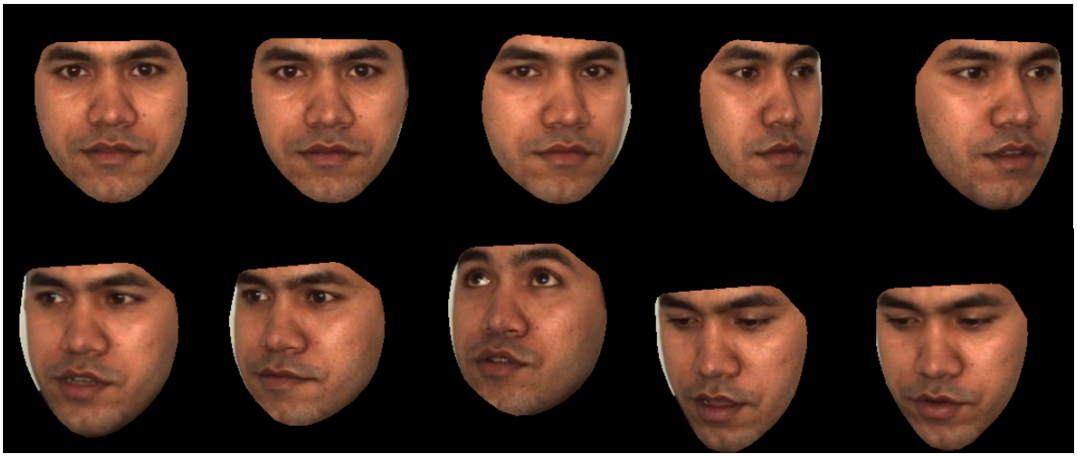


FIGURE 4.8 Texture is segmented from the face image sequence using shape model fitting. Each shot contains shape and texture information.

$$\alpha + \beta + \gamma = 1 \tag{4.20}$$

For a given point (x, y), the values of α, β and γ are given by:

$$\alpha = 1 - (\beta + \gamma) \tag{4.21}$$

$$\beta = \frac{yx_3 - x_1y - x_3y_1 - y_3x + x_1y_3 + xy_1}{-x_2y_3 + x_2y_1 + x_1y_3 + x_3y_1 - x_1y_2} \tag{4.22}$$

$$\gamma = \frac{xy_2 - xy_1 - x_1y_2 - x_2y + x_2y_1 + x_1y}{-x_2y_3 + x_2y_1 + x_1y_3 + x_3y_1 - x_1y_2} \tag{4.23}$$

Further, we perform LBP operator to extracted texture in order to obtain illumination invariance property of texture. A conventional LBP operator is applied to only those pixels which lie inside face area. This generates a code vector of length $256 = 2^8$ for gray scale images. We further calculate this vector for color images and obtain a code vector of size $256 * 3 = 768$. This high dimensional vector is processed through PCA to reduce the dimensionality and to use together with other feature set for classification.

4.4 Temporal Modeling

Local motion of the fiducial points is very crucial since it represents the facial deformations. This motion governs under facial muscles movements. We observe this motion by using optical

flow algorithm and displacement measures. For optical flow we use *Lucas-Kanade pyramidal algorithm* [Bou00] and relative motion of the individual points in consecutive frames.

A general *optical flow* algorithm approximates the local motion in a scene by using derivatives. There are certain limitations for the use of the optical flow algorithms which can however be avoided through the algorithm design. These limitations include:

- Motion is observed with global rigidity constraints. It is hard to find global deformations. In CKFE database , face do not exhibit global motion.
- Background motions are hard to model. We apply optical flow only in the detected face window.
- Occluded objects can produce false predictions.

For any given image, $I(x, y, t)$ denotes the pixel value at (x, y) location at time t. This pixel moves to the new position $I(x + \delta x, y + \delta y, t + \delta t)$ in time δt. Since $I(x, y, t)$ and $I(x + \delta x, y + \delta y, t + \delta t)$ are the same point, we can write

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t) \tag{4.24}$$

If $\delta x, \delta y$ and δt are very small indicating that the local motion is restricted in a small area, we can expand equation 4.24 using Taylor series expansion:

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t + H.O.T. \tag{4.25}$$

Since we assume that $\delta x, \delta y$ and δt are very small then we can safely neglect higher order terms (H.O.T.) and can write equation 4.25 as:

$$\frac{\partial I}{\partial x}v_x + \frac{\partial I}{\partial y}v_y + \frac{\partial I}{\partial t} = 0 \tag{4.26}$$

Where $v_x = \frac{\delta x}{\delta t}$ and $v_y = \frac{\delta y}{\delta t}$ are the x and y components of optical flow or image velocities. Further,

$$I_x = \frac{\partial I}{\partial x}, I_y = \frac{\partial I}{\partial y}, I_t = \frac{\partial I}{\partial t} \tag{4.27}$$

are image intensities derivatives at (x, y, t). Equation 4.26 can be re-written as:

$$(I_x, I_y).(v_x, v_y) = -I_t \tag{4.28}$$

or in compact form:

$$\bigtriangledown I.\overrightarrow{v} = -I_t \tag{4.29}$$

Where $I = (I_x, I_y)$ and $\overrightarrow{v} = (v_x, v_y)$ is the image velocity or optical flow at pixel (x, y) at time t. (Note: The terms are adopted from [BT05], for details, we refer to [BT05]).

Let I and J be the two gray scaled images of same size $w \times h$. Where w is the width and h is the height of the images. Consider an image point $u = [u_x\ u_y]^T$ on the first image I. The goal of feature tracking is to find the location $v = u + d = [u_x + d_x u_y + d_y]^T$ on the second image J such as $I(u)$ and $J(v)$ are similar. The vector $d = [d_x, d_y]^T$ is the image velocity and optical flow at x.

4.5 Spatiotemporal Multiple Feature (STMF)

We use different modules in hierarchy including face detection, shape model fitting, texture mapping and estimating optical flow-based parameters for feature vector extraction. The feature vector consists of the shape, texture and temporal variations, sufficient for considering local variations in shapes. All the subjects in the database are labeled for identification. Shape model from the training images is used for defining the reference shape in our experiments. This reference shape is calculated by finding the mean shape of the all shapes in the database.

An explicit 2D shape and appearance model is used to develop a baseline for feature extraction. This model comprises of 134 points which define the location of local face features like eyes, nose and lips in 2D space. The algorithm starts by locating a face in the image using haar-like features based on Viola and Jones face detector. An objective function is learned for fitting this model to the faces. After fitting the model to the example face image, texture information is extracted from the example image on a reference shape which is the mean of all the shapes in training data. Image texture is extracted using planar subdivisions of the reference and the example shapes. Texture warping between the subdivisions is performed using affine transformation. This image texture is now normalized both in the sense of shape and varying illuminations effects, making the image robust for shape and illumination. PCA is used to obtain the texture and shape parameters of the example image. This approach is similar to extracting AAM parameters. In addition to AAM parameters, temporal features of the facial changes are also calculated. Local motion of the feature points is observed using optical flow. These features are then used for classifiers for face recognition. A detailed process flow of our approach is shown in Figure 4.9. Our approach achieves real-time performance and provides robustness against facial expressions for real-world face recognition applications. In

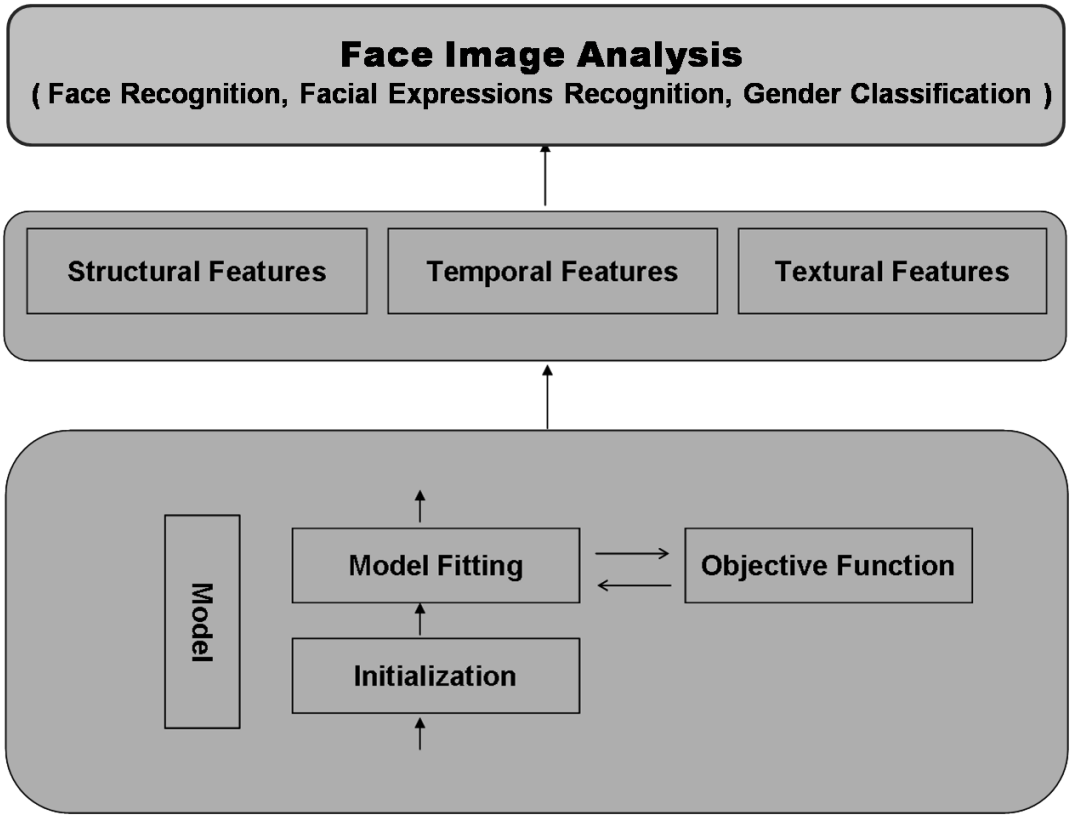


FIGURE 4.9 Model-based techniques split the challenge of image interpretation into computationally independent modules.

general, this computer vision task comprises of various modules coupled together and working sequentially by exploiting model-based techniques.

Representative features using model-based techniques consist of different modules: the model definition, the initialization algorithm, learning the objective function, the fitting algorithm, texture warping, motion features extraction and parameter configuration to form a feature set. The shape model contains a parameter vector **p** that represents its configurations, such as position, orientation, scaling, and deformation. A model is mapped to the surface of an image via a set of feature points, a contour, a textured region, etc. Referring to [ETC98], deformable models are highly suitable for analyzing human faces with all their individual variations. Its parameters $p = (t_x, t_y, s, \theta, \mathbf{b})^T$ comprise of the translation in 2D, scaling factor, rotation, and a vector of deformation parameters $b = (b_{s,1}, \ldots, b_{s,m})^T$ respectively. The later component describes the configuration of the face, such as the opening of the mouth, roundness of the eyes, raising of the eyebrows. This is shown in Figure 4.4.

The objective function $f(I, p)$ yields a comparable value that specifies how accurately a pa-

rameterized model **p** describes an image I. It is also known as the likelihood, similarity, energy, cost, goodness, or quality function. Without losing generality, lower values are considered to denote a better model fit. Traditionally, objective functions are manually specified by first selecting a small number of simple image features, such as edges or corners, and then formulating mathematical calculation rules. Afterward, the appropriateness is subjectively determined by inspecting the result on example images and example model parameterizations. If the result is not satisfactory the function is tuned or redesigned from scratch. This heuristic approach relies on the designer's intuition about a good measure of fitness. Earlier works [WSTR06] show that this methodology is erroneous and tedious.

To avoid these drawbacks, we used an approach that learns the objective function from annotated example images, proposed by [WSPR08]. It splits up the generation of the objective function into several tasks partly automated. This provides several benefits: firstly, automated steps replace the labor-intensive design of the objective function. Secondly, this approach is less error prone, because giving examples of good fit is much easier than explicitly specifying rules that need to cover all examples. Thirdly, this approach does not rely on expert knowledge and therefore it is generally applicable and not domain-dependent. The bottom line is that this approach yields more robust and accurate objective functions, which greatly facilitate the task of the fitting algorithm. The texture is extracted using affine transformation and gray values vector is formed as a raw feature. Gray values cannot be directly used because of their sensitivity against real world variations like changes in shape, illuminations, poses and expression. For this purpose, shape-free texture is also parameterized to represent the whole variation of the dataset with a small set of textural descriptors. Finally, the motion of each fiducial point is observed in corresponding frames. This represents the displacement of a particular point in the x and y coordinates. We again parameterize these vectors to find representative descriptor for motion representation.

4.6 Feature Extraction

Features are extracted using AAM approach with additional temporal parameters. We summarize step-by-step approach toward feature extraction.

4.6.1 Structural Features

For any input face image a model projected on the image plane and approximated with the fitting algorithm to find the best suitable structure of the given face. This structure defines

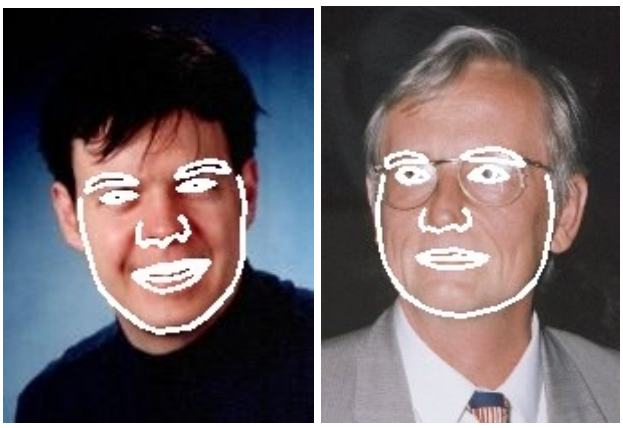


FIGURE 4.10 Examples of the model fitted to two random views.

the outer contour of the face and the position of the local facial features with 134 points. These points are coordinates in the image plane. We simply parametrize these points using equation 4.14. Figure 4.10 shows some example images describing shape of the face.

4.6.2 Textural Features

For various images of the same person different types of variations are required to be modeled. For example, shape deformations including both facial expression changes and pose variations along with the texture variations caused by illuminations. For this reason, different normalizations are required to be performed at this stage. At first, shape variation is required to be controlled in order to record the texture. This can be achieved by defining a reference shape for a particular object. In our case, this reference image is the mean shape, obtained by taking the mean of all the shapes of all persons in our training database. Figure 4.11 (bottom-left) shows the mean shape of the subject in consideration. Since the points distribution defines a convex hull of points in space so a planar subdivision is defined for the reference shape to map image texture. Delaunay triangulation is used to divide the shape into a set of different facets. Figure 4.11 shows the delaunay triangulations of the reference shape. Texture parameterization is performed using equation 4.18.

4.6.3 Temporal Features

Since facial expressions emerge from muscle activity, the motion of particular feature points within the face gives evidence about the facial expressions. These features further help the classifier to learn the motion activity. Real-time capability is important, and therefore, a small

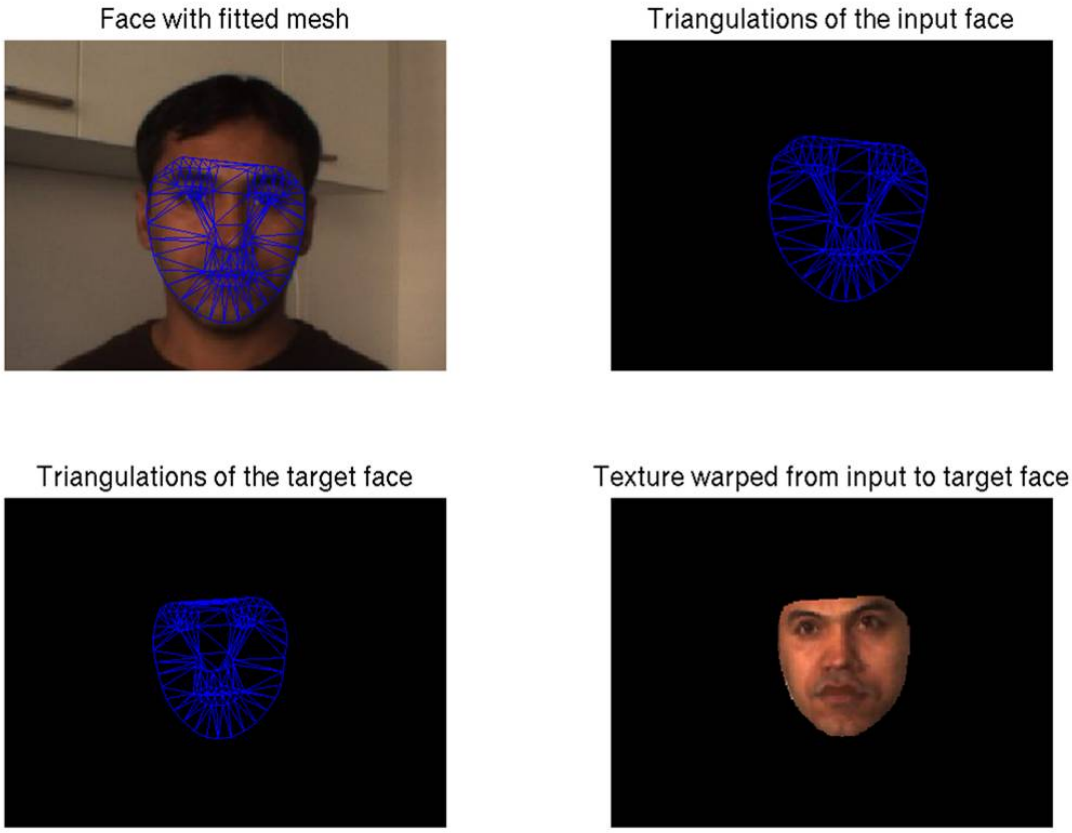


FIGURE 4.11 Texture extraction: An input face with a model fit (top left), triangular mesh of the input face (top right), triangular mesh of the target face (bottom left), texture warped from input face to target mesh (bottom right).

number of feature points are considered. The relative position of the model points is connected to the structural variations of the face model. Note that we do not specify these locations manually, because this assumes a good experience of the designer in analyzing facial expressions. We take benefit from the positions of the landmark points. We compute optical flow on these points using *Lukas Kanade pyramidal algorithm* [Bou00]. We parameterize again using PCA to extract temporal parameters.

If *t* is the velocity vector,

$$t = t_m + P_t b_t \tag{4.30}$$

Where temporal parameters b_t are computed using matrix of eigenvectors P_t and mean velocity vectors t_m.

The overall feature vector then becomes:

$$u = (b_{s,1}, \ldots, b_{s,m}, b_{g,1}, \ldots, b_{g,n}, b_{t,1}, \ldots, b_{t,p}) \tag{4.31}$$

Where b_s, b_t and b_g are shape, temporal and textural parameters respectively. Where m, n and p are the number of parameters retained from PCA space.

4.6.4 Feature Invariance

Since feature vector is formed with different types of parameters each of them has specific characteristics for face representation. Shape parameters describe the structure of the face which depends upon robustness of face detector. Since face detection algorithm is capable to work in the presence of slight in-plane and out of plane rotations and scaling hence the model can be fitted to face image within the presence of these variations. On the other hand, shape-free texture accounts for expressions invariance since we neutralize expressions during image warping. However, the final feature set has the ability to represent person identity, facial expressions and gender. Because facial expressions information is stored in shape and temporal parameters. Person identity information is mainly stored in textural parameters. Therefore, the extracted features not only provide the strong representation for the face images but also provide invariance against facial expressions. The experimental results are shown in section 4.7 in detail.

4.7 Experimental Evaluations

We perform extensive experiments on standard face databases and compare the results with the others approaches. Different types of experiments have been performed with 2D appearance models. We study face recognition, facial expressions, gender classification and age estimation. The goal of these experiments is to show the sufficiency of the feature set to deal with different face image classifications.

4.7.1 Model-based Segmentation

Robust facial feature localization has been one of the challenging issue. As explained in chapter 3 that minimum three points are required to normalize an input face image to a standard template. However this three point affine transformation is not robust and sufficient to normalize a face. For this reason a compensation to dislocation of facial features is performed using

FIGURE 4.12 Eigenfaces segmented using 2D model fitting. These eigenfaces corresponds to texture modes of the model.

clustering. Since the face models define the face boundary and facial features in more detail, they can suitably be used to extract texture from the region of interest.

Traditional holistic approaches like PCA are sensitive to the misalignment however, efforts have been made by the researcher to normalize the face images for classification in unconstrained environments. A point distribution model defines 236 triangular patches in 2D. Affine transform in now performed patch-wise and each triangular region is warped to corresponding region in standard template. Each image is normalized using this approach however, it may suffer affine distortion in those patches which are rotated or lie on the face edges. A solution to this issue is proposed in chapter 5. Figure 4.12 shows faces warped to standard template and corresponding eigenfaces.

We use decision trees to classify with 10 fold cross validation. We experiment on a subset of CKFE database consisting of more than 4000 images of 62 different persons. We choose different number of eigenvalues and test the performance of the classifier. A recommended variance for CKFE database ranges from 90% to 97%. We truncate eigenvalues at 97% of variance level. In first experiment, we normalize face images to a standard template and segment the texture. The recognition rate achieved from segmented 2D face texture was recorded 92.93% on CKFE database . Figure 4.13 shows the true positive rate and false positive rate from our experiments. The results indicate that texture is useful for facial recognition and works satisfactorily even in the presence of facial expressions. It can be seen that the Bayesian network perform better than decision tree as shown in Figure 4.13, when we use only textural feature for face recognition. However experiments in the next section show that the decision tree performs better than Bayesian network if the feature vector is composed of multiple fea-

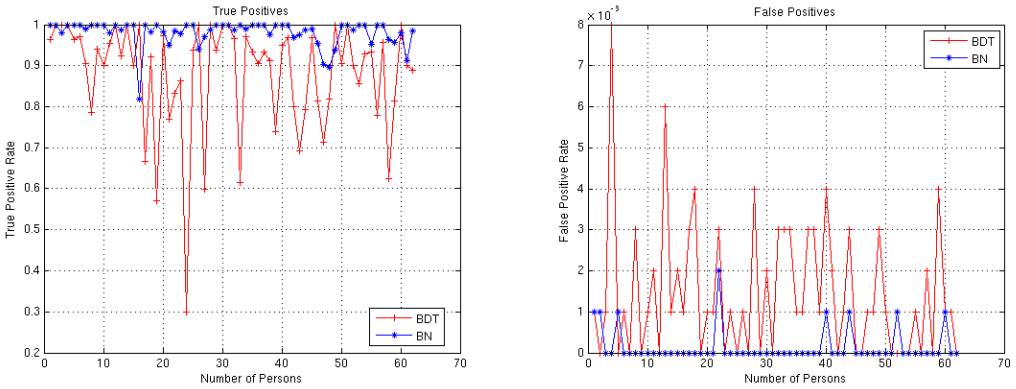


FIGURE 4.13 True positive and false positive values for face recognition on CKFE database with two different classifiers on 61 different persons.

tures. This fact is experimentally evaluated in section 2.3. A decision tree goes through each attribute and use tree pruning to purify until the last node. On the other hand, Bayesian network learn relation between the attributes and make a directed acyclic graph. Hence we prefer to use decision trees for experiments on STMF (For details refer to section 2.3 in Chapter 2).

In section 4.7.2 we further study effect of shape parameters to these texture parameters and get improved classification on the same database with same classifier settings. However, we also experiment in next section that if the same texture is used for facial expressions on the same database then it does not perform with high accuracy. The reason behind is the loss of useful shape information for facial expression classification. The texture in this case is shape free and only texture parameters are not sufficient to describe facial expression. The classification rate increases for facial expressions if we use shape and texture together as the feature. The classification rate is further improved if we observe this shape and texture with their temporal relation. This can be seen by the results of facial expressions in Table 4.2. Hence at this stage using a 2D face model, it can be clearly concluded that the shape, texture and temporal parameters are sufficient to represent facial recognition and facial expression. We do not put any necessity condition on the use of these features, however literature of face image analysis and our experiments show that such a compound feature set is highly recommended for face image analysis. In the next chapter 5 we also study these feature in detail and their invariance against real world challenges. A use of 3D face model further improves the results and robustness of our proposed approach.

4.7.2 STMF Evaluation

2D appearance models have been successfully used in four different experiments 1) face recognition in the presence of facial expressions, 2) facial expressions from image sequences, 3) gender classification and 4) age class estimation. These experiments have been performed on CKFE, MMI and FGNet databases (details about these databased is given in section 2.4) with the classifiers and classification methodology given in section 2.3.

Face Recognition: For experimental purposes, image sequences of 62 persons have been used which consists of overall 4060 images. A decision tree is trained as classifier in 22.99 sec. with splitting criteria and 10-fold cross validation. We used 1381 images for testing the recognition results and successfully recognized 1259 images. The recognition rate achieved was 91.17% in the presence of facial expressions. The higher accuracy corresponds to two major facts: 1) most of the faces are frontal in the images however there are slight in-plane rotations, 2) texture is warped to neutral face which creates neutralized facial expressions but on the other hand also create some affine distortions. Further, same set of the feature vectors are used to build a *Bayesian Networks* (BN) classifier. The classifier was built in 18.33 sec. using 10-fold cross validation. This produces a better accuracy of 98.69% of accurate recognition.

Facial Expressions Recognition: CKFE database does not contain natural facial expressions, but volunteers were asked to act. Furthermore, the image sequences are taken in controlled environment in a laboratory environment with predefined illumination conditions, solid background and frontal face views. The classifiers determine the facial expression for every image separately. Note that the information provided by the temporal features is still dependent on the previous images and therefore takes into account not only the information of one image but of all previous images in the sequence. We used 10-fold cross validation in both cases. Since each image sequence arises from neutral to peak of the given facial expressions then initial few images of the sequence contain neutral expression. These images are common in all sequence and cause higher confusion values in classification. In oder to overcome this problem, we filter out initial images where facial expressions are neutral. This is done by approximately ignoring first 8 to 10 images from each sequence. Another approach could also use neutral face as a seventh expression for classifier. However, we use the previous approach of ignoring the neutral expressions and classifying as six class problem.

We apply *support vector machines* (SVM) for pairwise classification of facial expressions. In a first experiment SVMs are trained with the temporal features only and achieved an accuracy of 79.5%. This result is improved to 92.2% by additionally providing the structural features. This is due to the fact that these features provide information about both, the single image and the image sequence. The best result is gained by providing full set of features and

Feature type	Shape + temporal	Shape + Texture	Shape + Texture + Temporal
Recog. rate	79.5%	92.2%	96.4%

TABLE 4.2 Facial expressions recognition using different combinations of the feature sets. It can be seen that the optimal performance is obtained using all three components. These results conform to concept of feature configuration given in Table 1.1.

the accuracy raised to 96.4%. This can be seen in Table 4.2. In order to create a classifier that is applicable in real-world environments as well, we train a classifier that is both, fast and capable of considering several class labels in the data. We apply decision trees for this task and provide them with the completely assembled data vector. The accuracy measured is 92.6%. Since its accuracy is only 4.1% below the best accuracy of SVM, this proves the high descriptive strength of the features provided.

Gender Classification: Classifying gender is a two class problem and hence we choose SVM as a classifier. Gender classification is performed in both supervised and unsupervised way. Shape parameters are obtained by fitting the model to the face images. One third of the extracted texture from all face images in the dataset is used to learn a PCA based subspace. The remaining texture is projected to this space and features vector is calculated as the weights in this eigenspace. The classifier is learned using 10-fold cross validation. We provide the classifier with two-third of the labeled data for training. The remaining one-third of the data which is not labeled is given to the classifier. A classification rate of 97.39% is achieved using decision tree as classifier with supervised learning.

Age Estimation: Further we estimate age using FG-NET database , detail refer section 2.4. This database contains 1002 images of 62 subjects with images of different ages ranging from 0 to 69 years. We divide the whole dataset in seven classes, where each class consists of ten years of age. Since the database consists of static images hence we experiment only with shape and textural component of the feature set. A classification rate of 49.70% is achieved with texture whereas the classification rate improved to 57.29% using support vector machine based classification. The *mean absolute error* (MAE) was 0.769. We compare our results with Hernandez et a. [RHCL10]. In this case we calculate MAE without segmenting the database in seven classes with ten years of age. We obtain an $MAE = 9.28$, better than multiple layer perceptrons (MLP) [LDC04] which is $MAE = 10.39$.

4.8 Summary and Conclusions

In this chapter we have implemented 2D human face model with the main focus on a point distribution model, called *Active Appearance Models* (AAMs) . The major contribution of this chapter is to extract a *Spatiotemporal Multiple Feature* (STMF) set. An STMF consists of three different components which are sufficient to study face image analysis. The individual role of the feature components has also been discussed. We explain different modules toward the development of a feature vector. For a given image, a face detector module finds an initial position of the face with a bounding box. A generic 2D face model is projected on this area using model parameters. We propose use of objective function based model fitting. This model is fitted to the given face image and structural information of this face is recorded using model parameters. We use this structural information to extract texture from the face area. An affine warp from the given face image to a reference face is utilized to store raw texture from face area. This texture is parametrized using PCA to obtain textural parameters. The relative position of the model points in the given image is observed in the next image of the sequence using optical flow. The output of the optical flow algorithm is the velocity vectors which are parameterized using PCA. From these three sequential modules, we obtain structural, textural and temporal parameters for the face image sequences. We form a unified feature set after parameter normalization and call this feature vector as STMF set. This feature vector is intensively experimented for different classification tasks. We conclude this chapter with following observations:

- Human face modeling provides a stronger and compact representation to study different aspects of human faces. Model based approaches are useful for interactive scenarios.
- Deformable models describe the detailed dynamics of the local facial features in faces and therefore can be used for facial expressions recognition.
- Temporal features play a significant role in analysis of facial expressions and hence they improve the classification rate as compared to conventional AAM. This can be seen in Table 4.2.
- The spatiotemporal feature extraction approach presented in this chapter is generic and can be applied to any deformable object.
- The proposed feature set can also roughly estimate age group by using structural and textural components of an STMF.

- The proposed multiple feature combination is useful for face recognition, facial expressions, gender classification and age estimation.
- 2D modeling is limited in dealing with varying poses because of the loss of depth information. A proposed solution to this problem is the use of 3D face model.
- Finally the feature configuration concept given in Table 1.1 in Chapter 1 has been proven experimentally that a combination of shape, texture and temporal parameters has high representation strength in different classifications.

Further extensions of this work is the use of a 3D model instead of 2D models which are restricted to image plane and lose depth information. In the next chapter, we study similar model but in 3D which further improves the results on these databases and verifies robustness of the proposed feature configuration. We pay special attention to structural and textural component of an STMF and prove its stability against varying poses and facial expressions in face recognition applications.

CHAPTER 5

3D Face Modeling

In the previous chapter, we have studied 2D face modeling in detail for extracting a spatiotemporal feature for image representation. The goal of this chapter is to improve the quality of the proposed *Spatiotemporal Multiple Feature* (STMF) and to pay a special attention to the textural component of this feature set. This chapter also studies the effects of perspective distortion on face images. Moreover, the improved feature set is invariant to varying poses and facial expressions on a face recognition system. In previous chapter, we have studied 2D *Active Appearance Models* (AAMs) in detail for face image analysis for face recognition, facial expressions recognition, gender classification and age estimation. In this chapter, we focus on 3D face modeling. The model is similar to a 2D appearance model and has a fewer number of vertices to define its structure. In 3D point cloud distribution models, point density plays a significant role in defining detailed non-rigid motions of the object. Coarser models, such as wireframe models used during our experiments are the conventional approach toward modeling an object. A wireframe in 3D represents a coarser mesh of vertices which are connected together and generate several *facets*. These facets are the non-overlapping regions on the surface of the object. The facets are generally triangular surfaces. Model based image interpretation requires context information from the image. Local facial features like eyes, nose, brows, lips etc. can efficiently be utilized to define the face model. For example, a task of face recognition requires the full area inside the face region including facial features and face boundary. For facial expressions, facial boundary might not be very significant but rather requires the exact location of the individual features and their motion.

The recommended use of 3D face modeling in HRI application is their ability to deal with the real world challenges. Human faces are seen in our routine life in a realistic and dynamic way conveying several information for interaction. A face in action can exhibit different head poses, meaningful communications through expressions, gender information, estimation of the age group, to some extent ethnic origin of the persons and especially feelings and behaviors

depending upon the context. 3D modeling can deal with illumination modeling, head poses and can generate a photorealistic model for computer vision analysis and synthesis of the faces. This property helps the machines to realize previously unseen views of a person. The goal of this chapter is to develop such kind of a photorealistic model to extract STMF in daily life scenarios where humans rely mostly on face to face interaction and interpret gender, identity, facial behavior and age of the other persons at a very first glance. We also discuss briefly some outstanding challenges like head poses and facial expressions image synthesis. Due to the diversity of the application domain and optimization of relevant information extraction for computer vision applications, we propose to solve this problem using an interdisciplinary 3D face model. The model is built using computer vision and computer graphics tools with image processing techniques. In order to trade off between accuracy and efficiency, we choose a wireframe model which provides automatic face generation in near real time. The goal of this chapter is to provide a standalone and comprehensive framework to extract useful STMF from a 3D model. Such features due to their wide range of information and less computational power, find their applications in several advanced camera mounted technical systems. Although this chapter focuses on multi-feature extraction approach for human faces in interactive applications with intelligent systems, however the scope of this chapter is equally useful for researchers and industrial practitioner working in the modeling of 3D deformable objects. This chapter is mainly specified to human faces but can also be applied to other applications like medical imaging, industrial robot manipulation and action recognition.

5.1 Introduction

Human face modeling has been one of the challenging fields over the last few years. By the availability of better hardware, improved algorithms and demand in the commercial market for personalized hardware like notebooks, mobile phones etc., several commercially available systems can interpret face images in an efficient way. However these systems are somehow generally limited to only one specific application domain. For instance, face recognition systems focus on identifying the person by reducing the effect of facial expressions and normalizing the face to near to neutral expression. Similarly facial poses are synthesized to frontal views for classification. As a general rule of thumb, such design approaches try to isolate the sources of variations and enable the system for one particular application. This approach is not quite useful for advanced intelligent systems. Currently, cameras are becoming a useful tool in human life and are the vital constituent of most of the intelligent systems. Over the availability of advanced hardware, better computational power and GPUs, graphics tools are being

usefully embedded in the computer vision applications to enhance the system performance. This resulted in development of the systems with underlying 3D computer vision applications which provides even more details than the conventional 3D methods for object reconstruction, analysis and manipulation. Image textures on the other hands provide a wide range of information for object analysis. A well-realized graphic object provides detailed configuration of the objects in 3D. Such realization provides sufficient information about shape, pose, light source and textures in an image. Moreover these attributes could be synthesized over time to get detailed dynamics and improved realization with additional temporal information. In this regard, we present a technique to develop a unified set of features extracted from a 3D face model. These features are successfully used for higher level facial image interpretation in different application domains. These features are extracted with the help of a coarse 3D wireframe model. We also study currently outstanding issues in human face realization and information interpretation. This includes head pose, lighting conditions, facial expressions and real time rendering. The extracted features are made stable over these variations and are capable to be used in different applications. The structural hierarchy of this chapter leads toward a STMF extraction. We proceed step-by-step providing essential knowledge about the topic.

These applications not only apply to the face image analysis in challenging environments but also emphasize on insufficiency of the traditional approaches for face image analysis. For instance, traditional face recognition systems have the abilities to recognize the human using various techniques like feature based recognition, facial geometry for recognition, classifier design and model based methods [ZCPR03] but on the other hand similar features are not sufficient for gender recognition or facial expressions recognition. Models due to their wide range of information in a compact parameter descriptors provide a better solution. In this regard, model based approaches have been very successful over last few years. Currently the available models used by the researchers are deformable models, point distribution models, rigid models, morphable models and wireframe models [ZC05].

The remaining part of this chapter is mainly divided in four sections. In section 5.2, we study briefly other face models and compare that how our model is different from others. A detail about related work on 3D face modeling is already given in section 2.1.3.2. Section 5.3 describes the goal of this chapter and background of the design information. Section 5.4 describes the core of this chapter toward the feature extraction. The feature extraction approach is studied in detail for 3D structure, texture extraction and temporal parameters. Finally, we give a brief overview of the applications of the system in section 5.6.

5.2 Background

In the recent decade advancement in the field of camera technology, their mountability on mobiles and availability of high computational powers in personal computers have increased the demand of the user to extract more information from the images for higher applications. For instance, face detection in cameras, automatic smile-capture cameras, face recognition in notebooks and intelligent vending machines working with facial age and gender. The future of interactive technologies will be relying on the facial image analysis, especially socially inspired robots need to know about the behavior of interacting persons at a very first glance. In this manner robots can friendly interact with the humans in the very first meeting and can adapt themselves easily to the habits of interacting person. The applications of such system can further be extended to intelligent robotic nurses for patients and assistive robots for elders. In this section we study the background knowledge required to describe the role of 3D model based approaches in intelligent interaction. We subdivide this section in two parts. In the first part we briefly describe the face modeling in computer vision. The second part of this section contains a brief survey of the recent work and results by the other researchers in 3D face modeling. We address mainly state of the art approaches toward the development of the feature extraction system.

5.2.1 Overview of 3D Face Modeling

Before proceeding further, we provide sufficient background about face modeling by discussing different state-of-the-art approaches. This sub-section gives an overview about these models, their applications, advantages and challenges. Besides ASM, AAM and active contours (or snakes) studied in detail in section 2.1.3.2, 3D face models have attained popularity in gaming applications, interactive applications and avatar generation.

Wireframe models are similar to point distribution models but defined over 3D space. They are surface models consisting of different landmarks. Candide model series is an example of wireframe model. Candide-I was introduced by Rydk [Ryd87] consisting of 110 points. Three different versions of Candide model are available. Candide-II [Wel91] consists of 160 vertices and 238 triangles. Candide-III is the final modification of previous versions consisting of 113 points and 184 triangular surfaces [Ahl01]. The model is coarser than morphable models (Blanz V. & Vetter, 2003). It provides better control over facial features motion using *Facial Actions Coding System* (FACS) [EF78][EFH02]. We will study Candide-III model in more detail later in this chapter. Other 3D realistic models are photorealistic models [WH04] and 3D morphable models [BV03b]. A photorealistic model in [PZVkC] is developed using two

images, frontal and profile face views. The original image texture and the synthesized texture are projected to the 3D head with improved model for hair and ears. 3D morphable models are famous realistic models and use laser scanner data for generation of the model. They are detailed models because of the dense distribution of points describing the surface of the face. This chapter focuses to get comparable results as that of the 3D morphable models but with a coarser wireframe model. The results are compared both in the sense of better visualization and diversity in their applications.

5.3 Main Focus of the Chapter

The main goal of this chapter is to extract multiple features for different applications at the same time providing an overview of the model based approaches for face images and implement a simple and useful platform to develop such systems. We split our approach explicitly in problem statement and proposed solution.

5.3.1 Problem Statement: Face-at-a-Glance Scenario

In past few years, model based approaches have attained a huge attention of the research community owing to their compact and detailed description for the objects. Models describe a large size image in small set of descriptors. These few descriptors are called model parameters. Object modeling generally lies in context based image analysis and hence they narrow the search domain in an image and precisely look for the object of interest for which they are designed. This reduces false alarms in finding an object. We address the problem in which a robotic system is able to extract a common feature set automatically from face images and is capable to classify gender, person identity and facial behavior. In addition to that, this feature set can also classify between different face classes. In such applications an automatic and efficient feature extraction technique is required which can interpret any possible face information. Currently available systems lack this property. A major reason is that researchers recommend to isolate the sources of variations from the given data while focusing in a particular classification problem. For example, in face recognition application, many researchers normalize the face in order to remove facial expressions variations to improve face recognition results. So the extracted features do not contain facial expressions information. We address an idea to develop a unified feature set which is used for different applications like face recognition, facial expressions and behavior and gender classification. Since, humans can get this information at a very first glance, so we term this problem as face-at-a-glance problem in human robot interaction

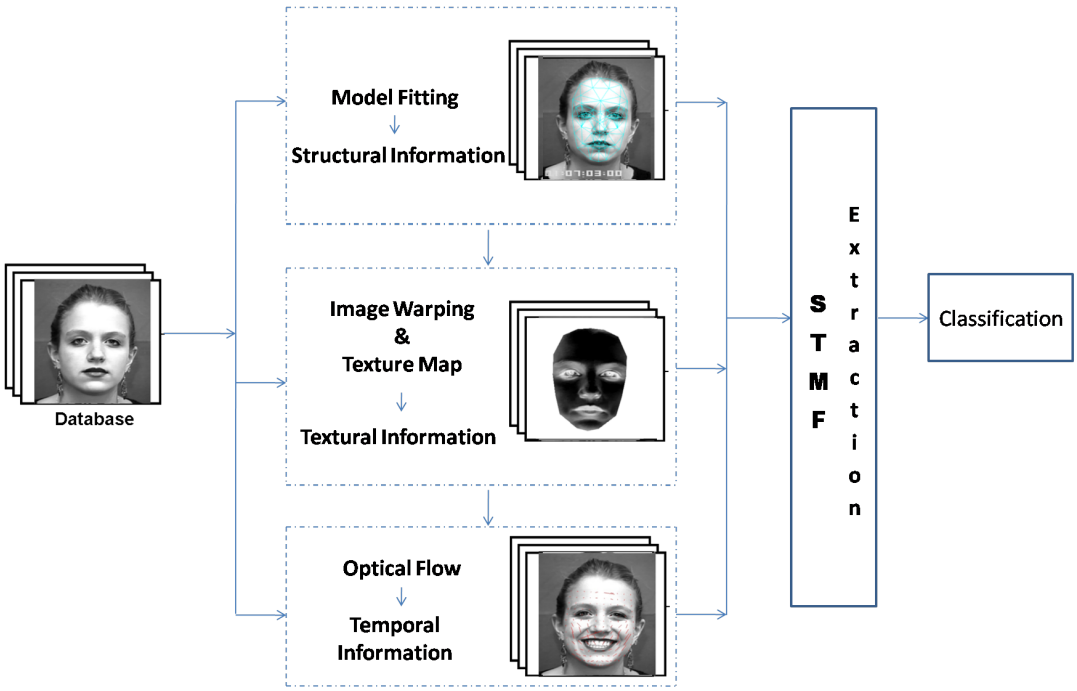


FIGURE 5.1 Overview of the different modules working towards STMF extraction.

domain. Further, in such scenarios faces are seen from different views under varying facial deformations and lightings. It is hard to find any solution in such case where a robot can find different information from the faces at the same time. Since 3D modeling can deal mostly with these challenges, we can propose a solution of finding useful features called STMF.

5.3.2 Proposed Solution: Spatiotemporal Multiple Features (STMF)

We propose multiple feature extraction as a recommended solution to this problem with some experimental evidence. Model parameters are obtained in an optimal way to maximize information from face region under various factors like facial pose, expressions, and illuminations. We use a wireframe 3D face model known as Candide-III for feature extraction, however any other model can also be applied. The model is fitted to the face image by learning robust displacement experts discussed later in detail in section 5.4.1. Model fitting provides the optimal set of parameters describing structure of the face in the given image. We contribute mainly in texture extraction and realization to obtain different types of texture features. These features show their strength in different classification problems. In case of 2D face modeling, texture information is mapped from the example image to a reference shape which is the mean shape of all the shapes available in the database. However the texture is not well defined on face

edges where the triangles are tilted. In order to avoid distortions in each triangular patch, we apply transformation of texture from image plane to frontal texture patches. This is achieved by comparatively applying affine and perspective transformation (details in section 5.4.5). With the best of two we choose perspective transformation. This undistorted texture is stored as a texture map which is an image with blocks where each block represents a triangle. The shape parameters additionally contain action unit activation levels. In addition to shape and texture parameters, temporal features of the facial changes are also calculated. Local motion of the feature points is observed using optical flow. We use reduced descriptors by trading of between accuracy and run time performance. These features are then used for classification. Our approach achieves real-time performance and provides robustness against facial expressions in real-world scenarios. Currently the system finds the pose information implicitly in structural parameters whereas illuminations changes are dealt in appearance parameters. This computer vision task comprises of various phases shown in Figure 5.1 for which it exploits model-based techniques that accurately localize facial features, seamlessly track them through image sequences, and finally infer facial features.

5.4 STMF Extraction

In this section we explain our approach in different modules including shape model fitting, image warping to a texture map, texture extraction, varying poses and expressions normalizations and finally synthesizing the image to extract model parameters. The proposed framework initializes with a coarse localization of the face image. If the robot is unable to find a person then cameras keep on finding a person unless they locate a face. Any coarse localization algorithm can be used for this purpose which can be refined in later stages of the system. We initialize by applying the algorithm of Viola et al [VJ04] to roughly detect the face position within the image. If the face is detected falsely in the image then case model fitting approach diverges and losses the control of the face position in the image in few frames. In this case, algorithm reinitializes itself for face search. In this section, we explain step-by-step our method toward STMF extraction and their fusion.

5.4.1 Model Fitting

For 3D face modeling we use a slightly different approach as compared to the objective functions, which was discussed in previous chapter. We use displacement experts using multi-band images. This approach is used by Mayer et al and for further details, we refer to [MR10].

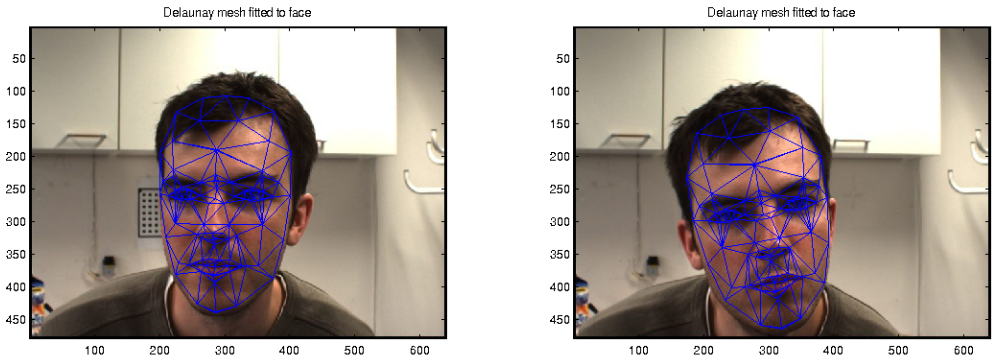


FIGURE 5.2 Fitting results with 3D wireframe model.

Model updates itself and adjustment to the face images are performed by using displacement experts at every individual points of the model.

Initial localization of the face in an image is used for model fitting. If the detected face window is located near to real face in the image then the model is capable to quickly adapt to the face. In the other case it needs some images to adapt to the model. If the face detector fails due to any reason, the system stops proceeding unless it finds an image with a face. A realistic face model relies strongly on model fitting. We opt wireframe model to fit to the face image on contrary to dense point cloud. The reason behind this is the efficiency and robustness of the fitting algorithm. The face detector localizes the face region within a bounding box. The Probability of the skin color is computed and local facial features like eyes and lip are obtained using their position and a color mask is generated. This approach is less prone to errors because of a better quality of the annotated images which are provided to the system for training. Further, it is less laborious because the design phase uses automated learning. As compared to objective function based approach discussed in previous chapter (details in section 4.2.2), displacement experts significantly increase accuracy of model fitting. Further, these objective functions are computationally expensive and prone to error due to local noise. A major issue in the use of displacement experts is the constraints on the local displacement Δp. This local displacement is restricted to the line perpendicular to the edge where the point is lying. If the value of Δp is large then fitting might diverge leading to false positives. This problem is treated using a near accurate facial features localization.

For a given image I, if the initial fitting parameters for this image are given by $\mathbf{p}$ then the parameter update $\Delta\mathbf{p}$ is required to be calculated using:

$$\mathbf{p}_I = \mathbf{p} + \Delta\mathbf{p} \tag{5.1}$$

Where $\mathbf{p}_I$ is the update parameters for the image. Ideally $\mathbf{p}_I = \mathbf{p}$. $\Delta\mathbf{p}$ is calculated from image contents and the given model parameters using the optimal solution of the following equation:

$$\Delta\mathbf{p} = g(I; \mathbf{p}) \tag{5.2}$$

Where $g(I; \mathbf{p})$ is the function of model parameters and features calculated from given image, similar to the objective function in section 4.2.2. This function is calculated from *multi-band images*. Multi-band images contain various information about the single feature in one band image. For example, Cootes et al [CT01] use two feature band images which represent the edges in two directions. The model parameters corresponds to relative positions of the facial features in the face image and hence the fitting algorithm is benefited from the exact localization of these features. Mayer et al [MR10] suggest the use parameters initialization from p_o:

$$\mathbf{p}_I = g(I; \mathbf{p}) + p_o \tag{5.3}$$

Figure 5.3 shows the three major steps toward the extraction of multi-band images. Displacement experts are extracted from four different features band. These bands include skin color image I^{skin}, brow image I^{brow}, lip image I^{lip} and retina image I^{retina}. Hence a multi-band image is formed with $I = \{I^{gray}, I^{skin}, I^{brow}, I^{lip}, I^{retina}\}$. In order to reduce the huge image data for the fitting algorithm, haar-like features with different sizes and orientations are used from the region around the facial features. The learning phase of the displacement experts is simplified by designing the experts for each parameter. Further, the feature search is reduced to the local regions. This is performed by considering the fact that parameters performing eyebrows motion do not effect the lip movements. Hence Δp is calculated from a set of parameter updates Δp_i and features set are extracted in the vicinity of the given feature:

$$\Delta p_i = g_i^l(I; \mathbf{p}_I^* + \Delta p_i) \tag{5.4}$$

Where $1 < i < n$, where $n = 85$ is the number of parameters representing structural features. Figure 5.2 shows model fitting results using this approach.

To extract structural features, the model parameters are exploited. The model configuration represents information about various facial features, such as lips, eye brows or eyes and therefore contributes to the extracted features. These structural features include information about the person's face structure that helps to determine person-specific information such as face deformations and facial expression generation. Furthermore, changes in these features indicate

FIGURE 5.3 Three steps to calculate multi-band images. For details refer to [MR10].

shape changes and therefore contribute to the classification of the facial expressions.

5.4.2 Structural Features

A Candide-III model consists of 113 vertices forming 184 non-overlapping triangles to define the surface geometry of the human face. As compared to AAM, this is a shape model whose geometry is controlled by a set of action units and animation units. The difference between these parameters is that the shape parameters control static deformation whereas animation parameters control facial expressions.

Any shape s can be written as a sum of mean shape $\overline{s}$ and a set of action units and shape units.

$$s\left(\alpha,\sigma\right)=\overline{s}+\phi_a\alpha+\phi_s\sigma \tag{5.5}$$

Where ϕ_a is the matrix of action unit vectors and ϕ_s is the matrix of shape vectors. Whereas α denotes action units parameters and σ denotes shape parameters. The scaling, rotation and translation of the model is described by

$$s(\alpha,\sigma,\pi)=mRs(\alpha,\sigma)+t \tag{5.6}$$

Where R and t are rotation and translation matrices respectively, m is the scaling factor and π contains six pose parameters plus a scaling factor. By changing the model parameters, it is possible to generate different views of the model to adapt it to the given face image. Figure 5.4 shows some of these deformations.

5.4.3 Facial Action Coding System (FACS)

Candied-III supports facial action coding system (FACS) by Ekmann and Freisen [EF78][EFH02]. Facial actions arise from the muscle movements inside the facial skins. They are direct reflection of these motions. Figure 5.4 shows a Candied-III model with global rotations and opened mouth face mesh generated using action AU13/15. Table 5.1 describes different action units,

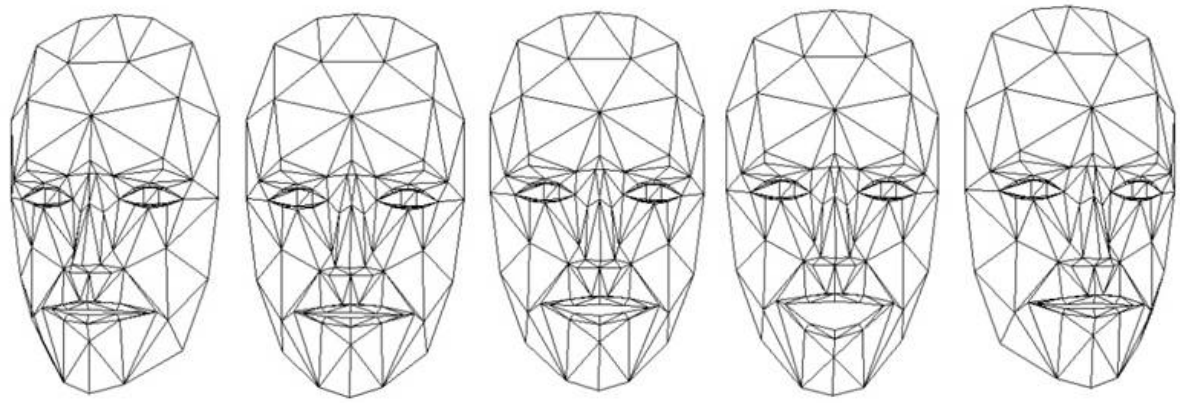


FIGURE 5.4 Structural variations governing under FACS principles and global rotations.

Action Unit	Action Unit Vector	Description	Muscle Motion
AU10	AUV0	Upper lip raiser	Levator labii superioris
AUV11	AU26/27	Jaw Drop	Masseter, relaxed Temporalis and internal Pterygoid
AUV2	AU20	Lip Strecher	Risorius w/ platysma
AUV3	AU4	Brow Lowerer	Corrugator supercilii, Depressor supercilii
AUV14	AU13/15	Lip corner depressor	Depressor anguli oris
AUV5	AU2	Outer brow raiser	Frontalis, pars lateralis
AUV6	AU42/43/44/45	Eyes closed	Relaxation of Levator palpebrae superioris; Orbicularis oculi, pars palpebralis
AUV7	AU7	Lid tightener	Orbicularis oculi, pars palpebralis
AUV8	AU9	Nose wrinkler	Levator labii superioris alaquae nasi
AUV9	AU23/24	Lip presser	Orbicularis oris
AUV10	AU5	Upper lid raiser	Levator palpebrae superioris

TABLE 5.1 Action units and action unit vectors with their visual effects in the generation of different facial expressions.

action unit vectors and corresponding muscle motions which are designed in Candide-III and Table 5.2 shows structural changes caused by different number of vertices.

5.4.4 Textural Mapping

In most of the 3D object modeling problems, objects are represented by separate structure defined by a mesh and a texture map which represents the appearance of the given structure. Heckbert [Hec86] define texture as a detailed pattern that can be repeated to tile on a plane, or a multi-dimensional image which can be mapped to a multi-dimensional surface. It is defined

No. of Vertices	Shape Variation
16	Head height
8	Eyebrows vertical position
36	Eyes vertical position
20	Eyes, width
24	Eyes, height
36	Eye separation distance
2	Cheeks z
6	Nose z-extension
17	Nose vertical position
3	Nose, pointing up
21	Mouth vertical position
14	Mouth width
36	Eyes vertical difference
2	Chin width

TABLE 5.2 Relationship between number of vertices and model deformations.

with the help of a mapping function which maps texture from the image to a surface. This process is divided in three major steps:

- Defining the texture map with texture coordinates (u, v).
- Mapping the texture from a texture map to 3D surface (x, y, z) of the object.
- Project the 3D object to screen coordinates (x_s, y_s) to display the object.

Figure 5.6 shows this process in detail. Texture mapping is the process in which texture is so called glued on a 2D plane, or from image to another 2D plane or 3D surface. This gluing of texture is performed with several sophisticated ways like decal modes, replace mode, modulate mode, blending etc. This depends upon the renderer that how it is defined to map the texture. We use rendering approach instead of rasterization, where texture and geometrical coordinates are used to render the objects. A single unit of texture map is called *texel* similar to a pixel in an image. As a texture map can also be a simple image, but a texel may not necessarily be a pixel. For example, if we want to paint a brick wall, it is required to paint each brick

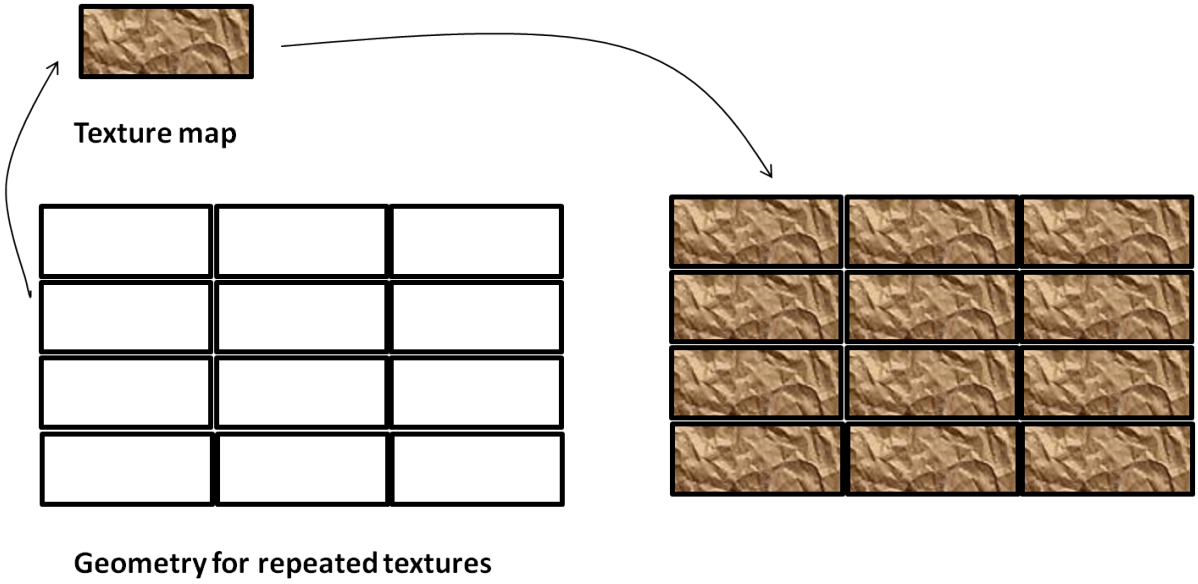


FIGURE 5.5 A single texture unit is pasted repeatedly on a brick wall pattern.

with a same texture, however properly defining that how a single brick may look at the edges. The final rendered wall should look realistic so that each brick could be distinguished from its neighboring bricks. In this case, the brick texture might be treated as a texel which consists of more than one pixel. Figure 5.5 shows a texture map, geometry and a rendered object. Texture coordinates are calculated which represent indices in an indexed texture map or direct coordinates in 2D map. These coordinates are associated with the vertices of the mesh defining the object.

Generally, four steps are involved in texture mapping, we use *OpenGL* for this purpose:

- **Creating a texture object:** In general, a texture is a 2D object however it can also be a 1D object. The data describing the texture consists of four components which are R, G, B and A values. Where A represents transparency level and is useful for blending.

- **Define texture mapping mode:** This step defines that how the texture is painted on the objects. This can be a decal mode, which simply maps texture to the surface of the object, replace mode, modulate mode etc. During this step, we define how texture will be mapped, either using shading and illumination models or not.

- **Enable texture mapping:** *Opengl* requires enabling of the texture using $GL\_TEXTURE\_1D$ or $GL\_TEXTURE\_2D$, which represents 1D texture enabling and 2D texture enabling respectively.

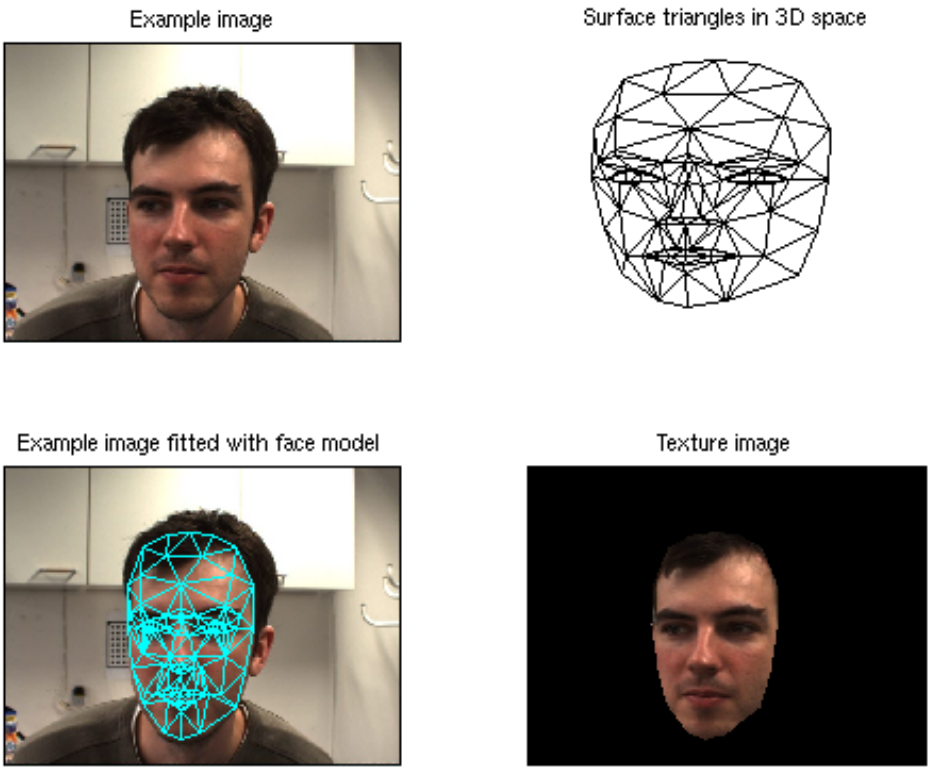


FIGURE 5.6 Texture mapping from image to 3D surface and displaying on screen.

- **Drawing the object:** Finally each vertex in the mesh of the object is associated with the texture coordinates and the mapping is performed according to the methods defined in above steps. Texture binding is performed to bind the texture.

In our case we have 184 texels in a texture map. We generate a texture map using general image transformations. The detailed process to extract a texture map is discussed in section 5.4.5. Once we extract a texture map, the face model is rendered. At this stage, we have texture coordinates and 3D vertices which define the structure of the face. Texture coordinates are simply generated by projecting the 3D vertices to the given image. For each image I, texture name is generated with *glGenTextures* and a given map is bound to this current texture name by using *glBindTexture*. A texture environment parameters are set to $GL\_TEXTURE\_ENV\_MODE$ which allows GL_DECAL in our case, used for decal mapping. Texturing parameters are set using $glTexParameteri$ with target as 2D texture and parameter name to $GL\_TEXTURE\_MAG\_FILTER$, which is the magnification func-

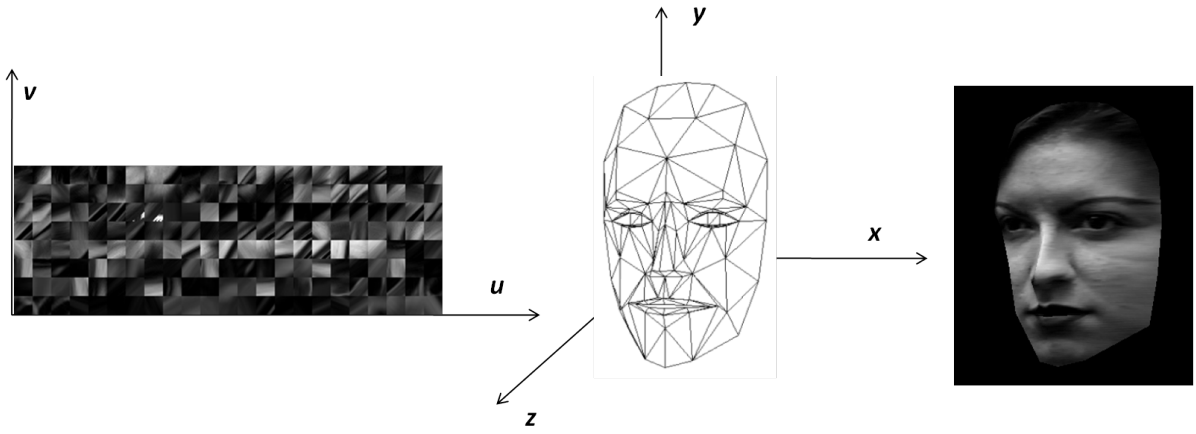


FIGURE 5.7 Texture is painted from a texture map to a face surface by using texture coordinates and 3D coordinated of the face model.

tion. The texture magnification function is used when the pixel being textured maps to an area less than or equal to one texture element. In our case, it sets the texture magnification function to GL_LINEAR. GL_LINEAR returns the weighted average of the four texture elements that are closest to the center of the pixel being textured. Similarly we use $GL\_TEXTURE\_MIN\_FILTER$ for texture parameters. The texture minifying function is used whenever the pixel being textured maps to an area greater than one texture element. Finally 2D texturing is enabled and the model is rendered triangle by triangle. Figure 5.7 shows the rendered model from the given blocks of texture maps.

5.4.5 Textural Features

The textural component of an STMF is calculated directly from the texture map instead of given image. This texture map is extracted after rectification of the texture at each triangular level. We apply PCA after texture mapping and parameterize the extracted texture. The approach is different to AAM texture extraction where the texture is extracted using affine transformation between the triangles using interpolation. We use undistorted texture in our case. Before we proceed to texture parameters, we explain here texture mapping in detail.

The robustness of textural parameters depends upon the quality of the input texture image. For example, the affine warping of the rendered triangle is not invariant to 3D rigid transformations of this triangle. It can be seen in Figure 5.9 (top row), that the affine warping works only if the triangle is not tilted with respect to the camera coordinate frame. If the triangle is tilted, as can be seen in Figure 5.9 (bottom row), the extracted texture is distorted. Since,

FIGURE 5.8 An example texture image, frontal triangle and tilted triangle with texture.

the 3D position of each triangle vertex as well as the camera parameters are known, we can determine the homogeneous mapping between the image plane and the texture coordinates. This mapping H is given by following formula:

$$H = K[r_1 r_2 - Rt] \tag{5.7}$$

Where K denotes the camera matrix, R denotes the rotation and t denotes the translation vector. It can easily be shown, that the formula above maps a 2D point of the texture image to the corresponding 2D point of the rendered image of the triangle. Since the 2D projection q of a general 3D point p in homogeneous coordinates can be written as follows:

$$q = K[R \; -Rt]p \tag{5.8}$$

It can be seen that each homogeneous 3D point lying on a plane with $z = 0$, i.e. $p = (x\, y\, 0\, 1)$ leads to above equation 5.8.

$$q = K \begin{bmatrix} r_1 & r_2 & r_3 & -Rt \end{bmatrix} \begin{bmatrix} x \\ y \\ 0 \\ 1 \end{bmatrix} = K. \begin{bmatrix} r_1 & r_2 & -Rt \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = Hp \tag{5.9}$$

with p being the homogeneous 2D point in texture coordinates. Since the camera parameters are known beforehand, the only values to be obtained are the rotation matrix *R* and the translation vector *t*. We use the upper triangle of a rectangular image for the texture values, and the triangles rarely fit this shape, we use an additional affine transformation *A*. The final homogeneous transformation *M* is the given by

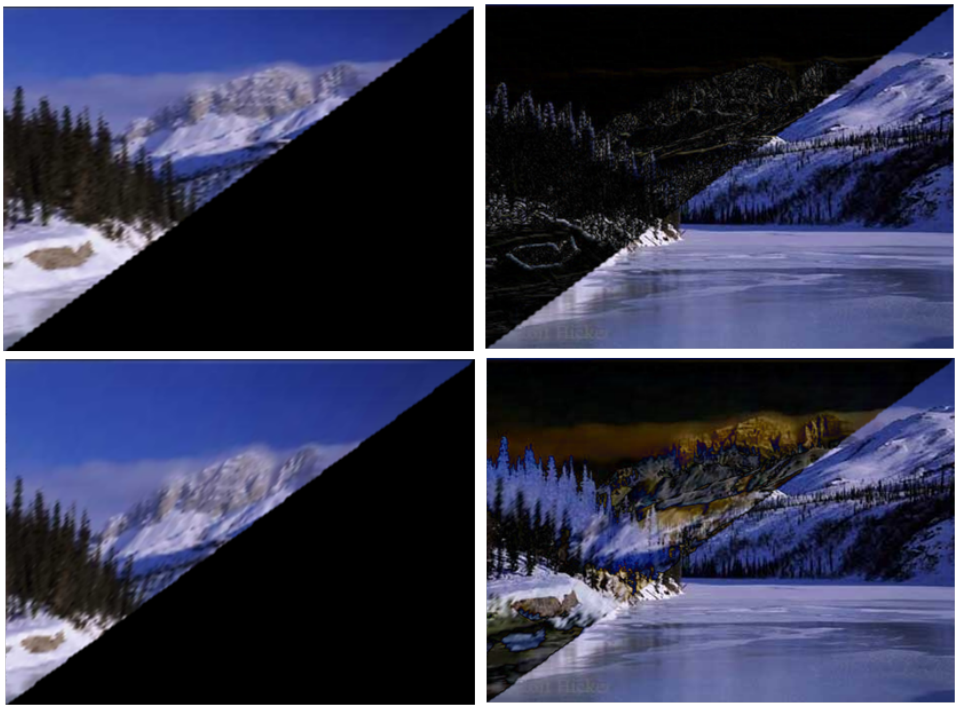


FIGURE 5.9 Affine transformation of the triangles given in Figure 5.8. (top row) shows the results for frontal triangle where texture projected to upper triangular area of the image, right image shows difference between this extracted texture with original texture, (bottom row) similar result for tilted triangle. It shows that in case of titled triangle texture is heavily distorted.

$$M = AK \begin{bmatrix} R & -Rt \end{bmatrix} = AH \tag{5.10}$$

It is determined in two steps. First we find H, by obtaining the rotation and translation of the triangle, by assuming that the initial triangle lies on the texture plane, the first vertex lies on the origin $(0, 0, 1)$ and the first edge lies on the x-axis. The affine transformation A is then calculated, so that the mapped triangle on the texture plane fits the upper triangle of the rectangular texture.

Figure 5.10 shows the same examples as for the affine warping. It can be seen that the extracted texture is not heavily distorted as in case of affine transformation. Besides the effects caused by the discretization of the image in pixels, the texture extraction is invariant against rigid transformations.

Once we have the shape information, the texture can be extracted after perspective correction from the image. Texture features are extracted using PCA, DCT and LBP. After comparative analysis of the recognition results of these features, it is found that PCA perform better

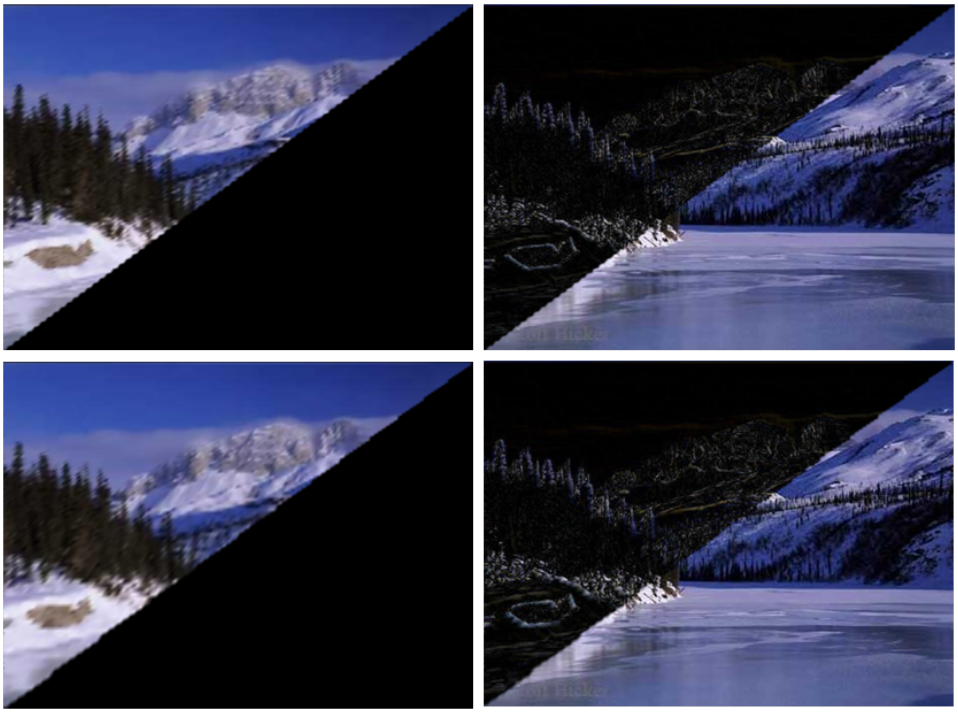


FIGURE 5.10 Perspective transformation of the triangles in Figure 5.8. (top row) results for frontal triangle where texture is projected to the upper triangular area of the image, right image shows the difference of this extracted texture with original texture, (bottom row) same result for tilted triangle. In this case extracted texture is much better and comparable to original texture which can be seen from the difference image.

than LBP and DCT. The reason behind these results is the texture representation is suitable for PCA rather than other two approaches. Since texture is realized in a detailed and better way, local descriptors can also perform better. We find energy based features which even outperforms PCA.

The extracted texture is parameterized using PCA by using mean texture $\overline{g}$ and matrix of eigenvectors P_g to obtain the parameter vector b_g.

$$g = \overline{g} + P_g b_g \tag{5.11}$$

Where g is a vector of pixel values from the face area.

5.4.6 Optimal Texture Representation

Each triangular patch represents meaningful texture which is stored in a square block of the texture map. A single unit of the texture map represents a triangular patch. We experiment with

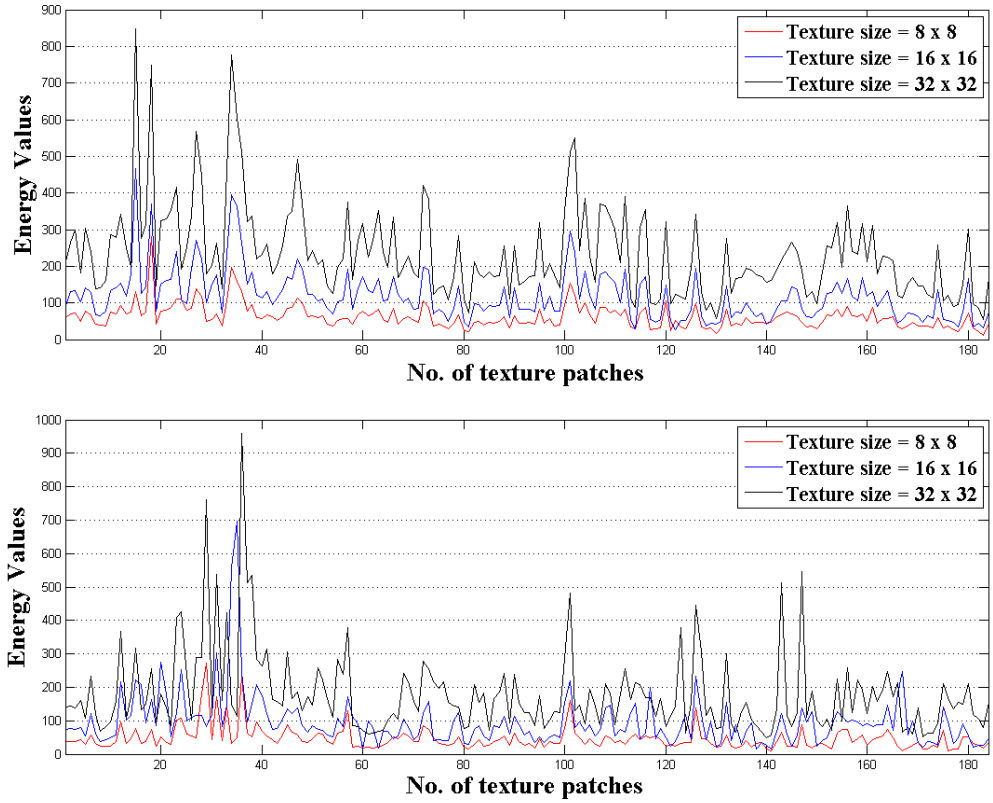


FIGURE 5.11 Energy spectrum of two randomly selected subjects from PIE database. Energy values for each patch is comparatively calculated and observed for three different texture sizes.

three different sizes of the texture blocks and choose an optimal size for our experimentation. These three block sizes include $2^3 \times 2^3$, $2^4 \times 2^4$ and $2^5 \times 2^5$. We calculate energy function from these texture maps of individual persons and observe the energy spectrum of the images in our database for each triangular patch. If N is the total number of images, and p_i be a texel value (which is equal to a single pixel value) in texture map, then we define energy function as:

$$E_j = \frac{1}{N} \sum_{i=1}^{N} (p_i - \overline{p}_j)^2 \tag{5.12}$$

Where $\overline{p}_j$ is the mean value of the pixels in j^{th} block, $j = 1 \ldots M$ and $M = 184$ is the number of blocks in a texture map. In addition to Equation 5.12, we also find variance energy by using PCA for each block and observe the energy spectrum. The variation within the given block has similar behavior for two kinds of energy functions except a slight variation in the energy values. Figure 5.11 shows the energy values for two different subjects randomly chosen

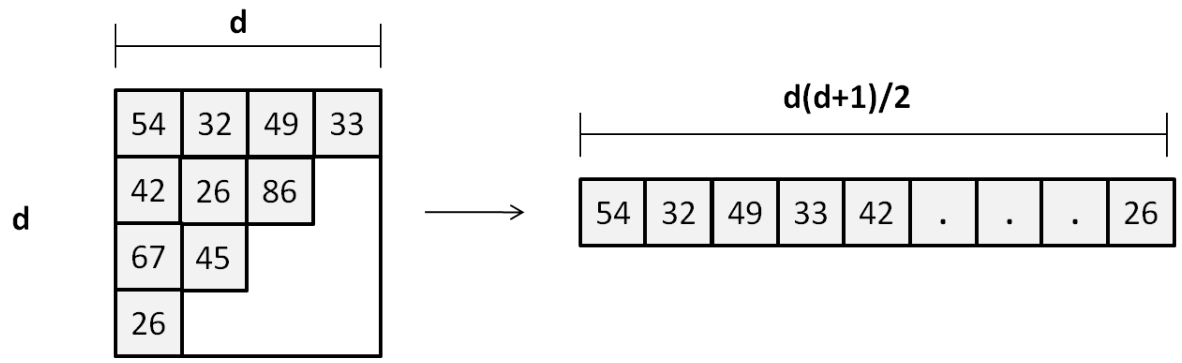


FIGURE 5.12 Texture from each triangular patch is stored as upper triangle of the texture block in texture map. A raw feature vector is obtained by concatenating the pixel values from each block.

from our experiments. It can be seen from Figure 5.11 that behavior of the textural components is similar between different texture sizes. The size of the raw feature vector extracted directly from texture map increases exponentially with the increase of texture block size. If $d \times d$ is the size of the block, then the length of the raw feature vector is $\frac{d(d+1)}{2}$. This vector length calculation depends upon how texture is stored in the texture map. This can be seen in Figure 5.12. We store each triangular patch from the face surface to upper triangle of the texture block. The size of raw feature vector extracted for $d = 2^3$, $d = 2^4$ and $d = 2^5$ is 6624, 25024 and 97152 respectively. Any higher value will exponentially increase the raw vector without any improvement in the texture energy. We do not consider higher values due to increase in vector length. The overall recognition rate produced by different texture sizes from eight randomly selected subjects with 2145 images from PIE database is shown in Figure 5.13. The results are obtained using decision trees and Bayesian networks for classification. By trading off between the performance and size of the feature vectors, we choose texture block size to 16×16 during our experiments.

5.4.7 Face Synthesis and Texture Extraction

We use texture patches which are obtained after removing perspective distortions for each triangle and store the texture in a texture map. Our texture map consists of blocks where each block represents a triangle. Undistorted texture is stored as upper left triangle (as shown in Figure 5.15). This texture map can be used for two different purposes at this stage. Firstly, we can synthesize different views of a person at this level. Since we already have shape and texture information in the form of parameters, it is quite simple to synthesize different views of a person. By changing rotation, scaling and translation vector we can generate some global

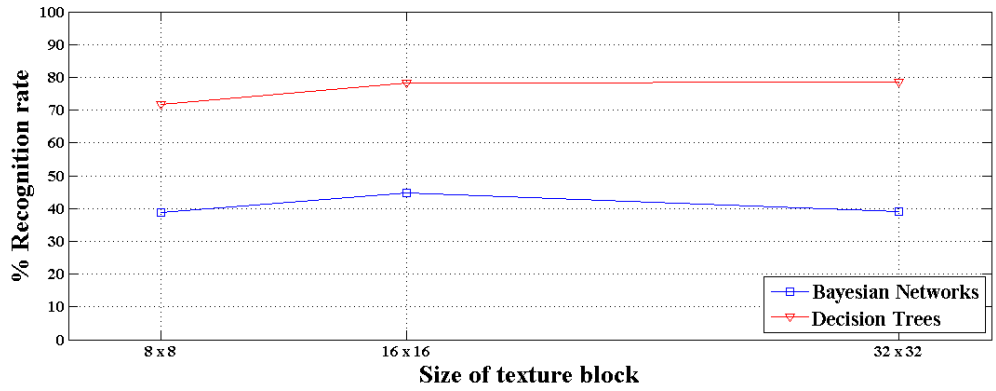


FIGURE 5.13 Comparison over eight random subjects from the database with three different sizes of texture blocks. Recognition rate slightly improved as texture size is increased however causes a high increase on the length of raw feature vector. We compromise on texture block of size 16×16.

motion. This can be seen in Figure 5.14 where half profile view of an example face (Figure 5.14, last row) is synthesized for rotation across vertical axis. Further examples are shown in Appendix-A.

Texture map is stored for each view seen and further views can be synthesized. In order to find textural parameters, we use equation 5.11.

Since Candide-III supports FACS, it is quite useful to synthesize facial motions especially during training phase. This can help in learning person independent features and learning the faces with novel views. For example, laugh is synthesized from the neutral face by mainly varying action unit *AU13/15* as shown in Figure 5.16.

5.4.8 Temporal Features

Continuous structural variations in a face generate different facial expressions. In order to record facial deformations, we observe temporal behavior of various facial components using optical flow. Facial expressions arise from a set of action units defined in Table 5.1 which are caused by muscular movement behind face skin. Since our model consists of fiducial points which correspond to different facial features, hence it is quite useful to observe the motion of these points in an image sequence to record temporal changes. We apply *Lukas-Kanade pyramidal* implementation on these points and calculate the velocity vectors in corresponding images. These velocity vectors are parametrized using PCA as explain in previous chapter (section 4.4).

If t is the velocity vector,

FIGURE 5.14 Model Fitting to examples image and texture projection on 3D surface after perspective correction (©images from PIE-database).

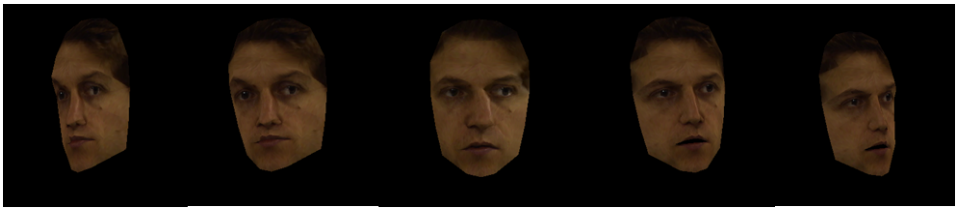


FIGURE 5.15 Synthesized poses from Figure 5.14 (bottom row) by changing global rotation and FACS units.

FIGURE 5.16 Texture Map and synthesized novel view from Figure 5.14. The storage pattern of each texture block is same and is given in Figure 5.12.

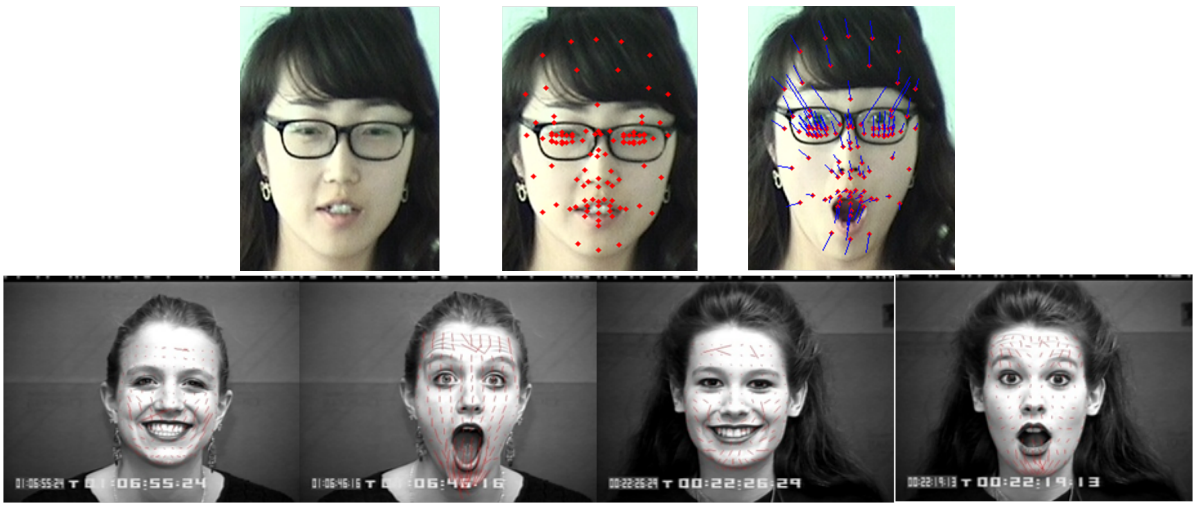


FIGURE 5.17 Optical flow of different points using Locas-Kanade pyramidal algorithm [Bou00][Wim07].

$$t = t_m + P_t b_t \tag{5.13}$$

Where temporal parameters b_t are computed using matrix of eigenvectors P_t and mean velocity vectors t_m. Figure 5.17 shows motion pattern examples images.

5.4.9 Feature Fusion

We combine all extracted features into a single feature vector. Single image information is considered by the structural and textural features whereas image sequence information is considered by the temporal features. The overall feature vector becomes:

$$u = (b_{s,1}, \ldots, b_{s,m}, b_{g,1}, \ldots, b_{g,n}, b_{t,1}, \ldots, b_{t,p}) \tag{5.14}$$

Where b_s, b_g and b_t are shape, textural and temporal parameters respectively with m, n and p being the number of parameters obtained from subspace in each case. Equation 5.14 is called *Spatiotemporal Multiple Feature* (STMF) . Since features arise from different sources, it is not quite obvious to fuse them together to get a feature set. This can cause the dominance of the features with higher values and ones with low values are ignored [FCGH08]. We use simple scaling of the features in $[0, 1]$. However, any suitable method for feature fusion can be applied here. We extract 85 structural features, 74 textural features and 12 temporal features textural parameters to form a combined feature vector for each image. These features are

then used for *decision tree* (DT) and *Bayesian network* (BN) for different classifications. The face feature vector consists of the shape, texture and temporal variations, which sufficiently defines global and local variations of the face. All the subjects in the database are labeled for classification.

5.5 Experimental Evaluations

The feature set obtained in this chapter consists of similar configuration as given in last chapter however modification is performed in structural and textural parameters. The structural parameters are 3D approximation from a single image. This structure represents rigid and non-rigid deformation of the face. The texture is calculated using homography where each triangular patch is stored after perspective texture improvements. Different types of experiments have been performed using this feature set. We study face recognition, facial expressions and gender classification. The goal of the experiments is to show the sufficiency of the feature set to deal with different face image classifications while showing robustness against poses and facial expressions at the same time for face recognition system.

5.5.1 Model-based Segmentation

The model used in our experiments covers the whole face area except ears and hair whereas 2D model does not cover forehead. We segment the face images to a neutralized standard face template by using texture warping. This extracted texture is used for face recognition in a similar way which is used for 2D models in section 4.7.1. A recognition rate of 97.75% was obtained on these image sequences in case of 3D model. We train our classifier with same number of images and classifier with same specification and achieved a recognition rate of 92.93% for 2D model (for detail refer to section 4.7.1). Figure 5.18 shows segmented images from 2D and 3D face model and Table 5.3 shows true positive rate (TPR) and false positive rate (FPR) for these models.

5.5.2 Texture Rectification

In order to study perspective effect on face images, we experiment mainly on CMU PIE database [TBB02] for face recognition and verify this fact on FG-NET database [fgn] for age estimation. There are two sessions of CMU PIE database captured from 1) October 2000 to November 2000 and 2) November 2000 to December 2000. This database contains pose, illuminations and facial expression variations. Since our algorithm starts with Viola and Jones

FIGURE 5.18 Face segmentation using 2D model (left), and 3D model (right).

face detector [VJ04], hence we consider only those images where face detector results are positive. In this filtration we obtain 3578 images where faces are successfully detected. The images with high pose variations (profile poses), darkness effects and illuminations are filtered out in this step. We obtain images with frontal, half profile in both directions, looking upward and looking downward faces. Texture is extracted from each image by using method described in section 5.4.5. We obtain 3578 vectors of 25024 length. For dimensionality reduction, we use 40% of the data randomly selected from these raw features to learn PCA based subspace. For all texture sizes, we retain 97% of the covariance by choosing among the eigenvalues. The remaining data is projected on this space to obtain parameters which serve as feature vectors. The size of the parameter vector for three different textures is almost equal. This vector length is 188, 185 and 186 respectively for texture sizes of 8×8, 16×16 and 32×32.

For classification purpose, we apply decision tree. However, other classifiers can also be applied depending upon the application (*Bayesian Networks* (BN) were also used with comparable results during experimentation (refer Figure 5.13)). We choose J48 decision tree with 10-fold cross validation algorithm for experimentation which uses tree pruning called subtree raising and recursively classifies until the last leave is pure. The parameters used in decision tree are: confidence factor C = 0.25, with minimum two number of instances per leaf and C4.5 approach for reduced error-pruning [WF05]. Face recognition rate under varying poses and facial expressions is given in Table 5.4. In order to verify the effect of perspective distortions, we further study age classification from all subjects of FG-NET database . This database consists of 1002 images of 62 subjects with age ranging from $0 - 69$ years. We divide the database in seven groups with 10 years band. The results are shown in Table 5.4. This table clearly shows

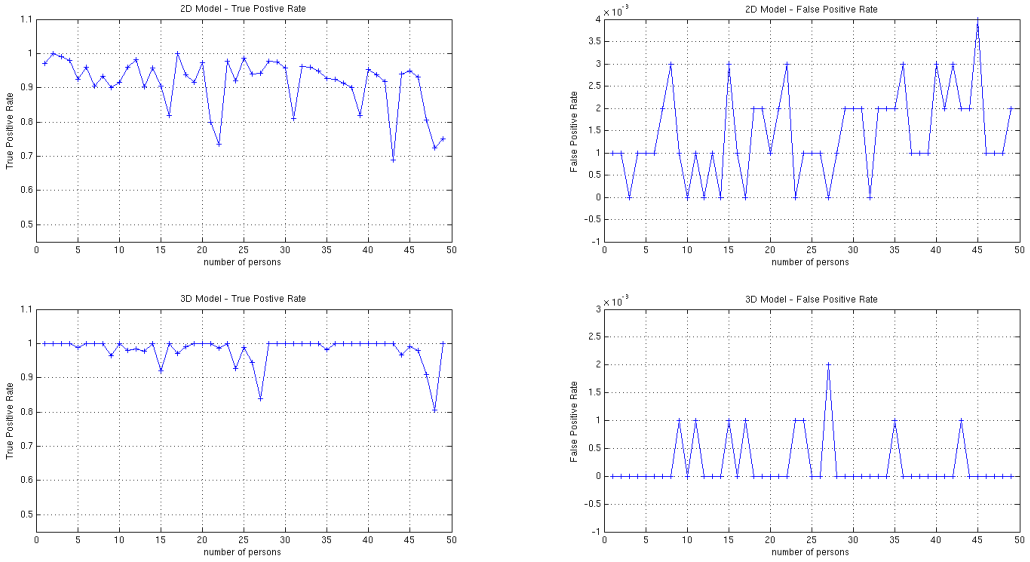


TABLE 5.3 True positive and false positive rate for face recognition on CKFE database . The face recognition results are obtained from 50 subjects of CKFE database in the presence of six different facial expressions. It can be seen that under same experimentation conditions, 3D face segmentation (bottom row) outperforms 2D face segmentation (upper row).

Database	**2D Texture parameters**	**Rectified Textural Parameters**	**AAM**	**3D Structural + Textural Parameters**
PIE	63.02%	**69.64%**	79.93%	**84.15%**
FG-NET	51.35%	**54.09%**	51.15%	**55.39%**

TABLE 5.4 Comparison of traditional AAM approach and rectified texture. The results are shown for textural parameters and combined structural and textural parameters.

that the texture extracted after detail texture map extraction and considering perspective effects on face images improves the classification rate as compared to state-of-the-art AAM.

5.5.3 3D Model-based Face Image Analysis

In order to experiment feature versatility we use two different classifiers with same feature set on three different applications: face recognition, facial expressions recognition and gender classification. Table 5.5 shows different recognition rates achieved during experimentations. This can be analyzed from *receiver operating characteristics* (ROC) curves. Figure 5.19 shows ROC curves for six different facial expressions. Since laugh and fear are often confused facial

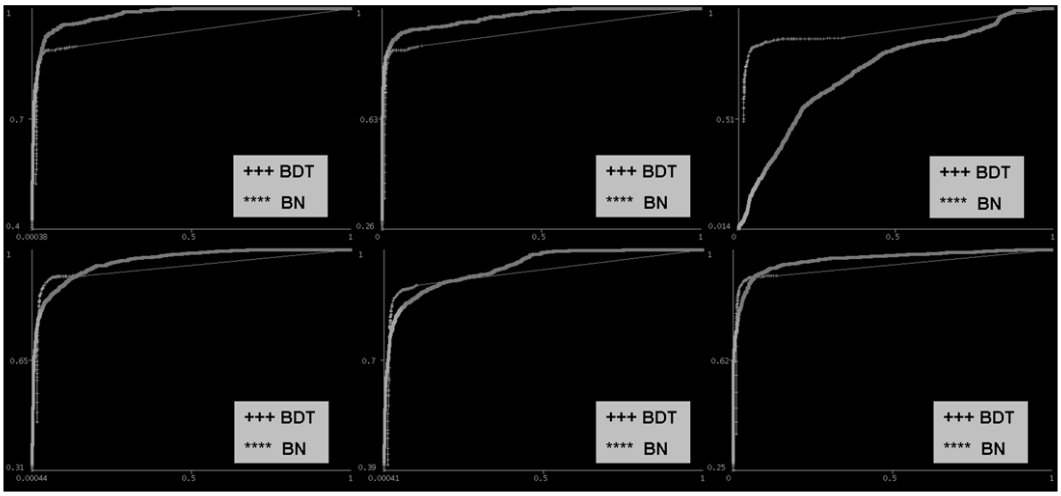


FIGURE 5.19 ROC curves for six different facial expressions. They indicate anger, disgust, fear (top row), laugh, sadness ans surprise (bottom row), for two different classifiers

	BDT	BN
Face Recognition	98.49%	90.66%
Facial Expressions Recognition	85.70%	80.57%
Gender Classification	99.08%	89.70%

TABLE 5.5 Three different classification results on CKFE database using decision tree and Bayesian network.

expressions, it can be analyzed from the curves that there exists some confusion between these two expressions for BN classifier, which is improved using BDT classifier. This shows the role of classifier for STMF. Figure 5.20 shows gender classification results for these classifiers.

Table 5.5 shows different recognition rates achieved during experimentations on CKFE database . In all three cases decision tree outperforms *Bayesian* networks. The choice of the classifier is explained in detail in section 2.3.

Since our feature set arises from different sources, so BDT is used for classification (details in section 2.3 in Chapter 2). We choose J48 decision tree algorithm for experimentation which uses tree pruning called subtree raising and recursively classifies until the last leave is pure. We use same configuration for classifier training during facial expressions and identity experiments. The parameters used in BDT are: confidence factor C = 0.25, with two minimum number of instances per leaf and C4.5 approach for reduced error pruning. We utilize Weka

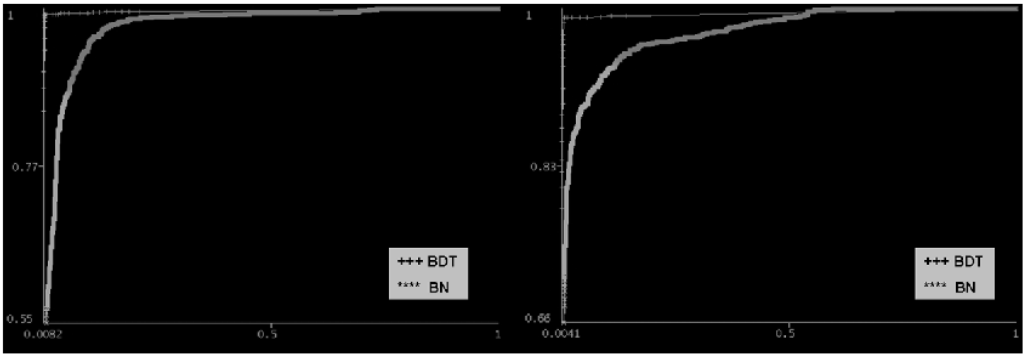


FIGURE 5.20 ROC curves for gender classification on two different classifiers, female (left), male (right).

[12] implementation for experiments. Figure 5.21 shows results of facial expression analysis for our laboratory capture database and CKFE database . We use person dependent expression classification for both databases. For both databases, we use all subjects in the database however do not consider neutral face as an additional expression. A neutral expression can be considered as a seventh expression. We filter neutral images from each sequence because our approach is image based instead of video based and neutral expression in each image sequence may cause more confusions among the other expressions. Our representative features show sufficient strength for facial recognition. We use same classifier for face recognition experiments with 10-fold cross validation. The overall facial recognition on CKFE database is 99.59% whereas on our laboratory captured database is 98.96%.

So far, we have experimented to prove that a multi-feature has ability to classify various facial attributes. In order to verify feature invariance against real world challenges, we perform experiments for face recognition in the presence of varying poses and facial expressions. We experiment on PIE database to test this feature. We again use decision tree as classifier with 10-fold cross validation on all images of the second subset of PIE database (collected from October to November). There are 9152 images in total however we get only 3578 images where face detection algorithm works and model parameters can be extracted. In this regard, we can say that the algorithm robustness depends upon face detection results. Table 5.6 shows the results with textural and combined parameters. In order to test the versatility and robustness of the proposed features, we experiment using different feature combination and come to the conclusion that 3D model based feature are suitable for real world applications.

Table 5.7 shows the comparison of our results with other approach for facial expressions

FIGURE 5.21 Confusion matrices for facial expressions recognition. Confusion matrix for five facial expressions including pain detection on our laboratory capture images from pan tilt zoom (PTZ) camera (left), Confusion matrix for CKFE database .

	LBP Gray	LBP Color	DCT Gray	DCT Color	PCA
Shape + Texture	78.56%	76.75%	78.51%	80.60%	83.06%
Texture	55.76%	56.46%	47.90%	46.81%	69.40%

TABLE 5.6 Performance of different features on PIE-database in face recognition application. The results show that PCA outperforms other features.

recognition and Table 5.8 shows the results for gender classification as compared to other approaches.

5.6 Applications

This work is intended for robots which are designed to work for elderly people in an assistive environment. This can help people to perform their activities safely and freely in the presence of an intelligent assistant. This kind of application is useful for personal robots [Sys]. By the rapid growth in the field of robotic technology it will be quite common to get a personal robot as like personal computer in past days. Besides this, such system can be applied to any scenario which are assisted by robots. For example, in medical imaging where the objects under treatment are deformable during surgeries. Further applications include action recognition, driver fatigue analysis, stress analysis and behavioral analysis of patients during treatment. The system takes input from a singe camera. The current system can be enhanced and configured according to its application with other sensory data. Due to its efficiency it can be integrated

Approach	% Accuracy
model based [MWER09]	87.1%
TAN [CSG+03]	83.3%
LBP + SVM [SGM09]	92.6%
IEBM [ASWG09]	92.9%
Fixed Jacobian [ASWG09]	89.6%
STMF + DT	**93.2%**

TABLE 5.7 Facial expressions recognition in comparison to different approaches given by [ASWG09].

Approach	Classification rate
Pixels + SVM + Fusion	88.5%
LBP + SVM + Fusion	92.1%
VLBP + SVM	84.5%
EVLBP + AdaBoost	84.6%
STMF+ BDT	**94.8%**

TABLE 5.8 Performance of gender classification results in comparison to different approaches in [HP09]

with other modalities like human voice, body language analysis, psychological analysis with fMRI.

5.7 Summary and Conclusions

This chapter thoroughly studies the development of a 3D face model from a single image using a coarser wireframe model describing the surface structure of the human face. This model supports deformation caused by FACS and hence provides an improved framework for multiple face image analysis applications. For a given image, a model is project to image plane. The projected model is a 2D projection of our 3D surface model. This projection is fitted to the face image using *displacement experts* approach. The fitted model provides structural information of the given face. This structural information is used to extract texture from the face area. Texture is extracted using two different approaches. In the first approach, we use image warping technique which was used in last chapter (refer to section 4.6.2). In second approach,

we generate texture map for each image and store undistorted texture patches in this texture map. This is implemented by using *homography*. The facial deformation is observed using motion of the model points. We study different textural features and find that the texture map representation is better than conventional texture warping approach. The realistic 3D model is generated from a single image and hence 3^{rd} dimension is approximated by using deformation parameters. A feature vector is constructed from structural, undistorted textures and temporal parameters which is finally used for different classification experiments. These experiments include face recognition, facial expression recognition, gender classification and human age estimation. The proposed feature set also provides robustness against facial expressions and poses for a face recognition system. Finally we conclude this chapter with following:

- This chapter provides a simple method to construct a 3D human face model by using computer vision and computer graphics approaches. We generate a 3D face model from a single image per face and successfully use it for face image analysis.
- 3D models due to their detailed representation of an object facilitate in design phase to deal with real world challenges like varying lights, head poses and facial dynamics. However these models are limited against high occlusions because of the model fitting issues and texture distortion.
- For point distribution models, effects of perspective distortions are generally ignored by the researchers in face image analysis because the distance of the camera from the face is very large as compared to the face size. However we prove from intensive experimentations that these effects are considerable and obtain a improved classification rate in face recognition and age estimation.
- Each triangular surface of the model is given equal weight in texture map, hence the triangles which are tilted on the face edges equally contribute to the feature vector. Therefore, the extracted feature set is found robust against varying poses for a face recognition system. This observation is obtained by comparing the results of conventional texture warping and texture map approach from our studies.
- In addition to standard databases, it is shown that this model can successfully be used for pain detection for patients or elderlies on a limited laboratory captured database. This property is due to the fact that FACS are is defined very well with the baseline model.

3D human face modeling come across with a few challenges during design and development phase. These challenges include efficiency, real time 3D reconstruction, light modeling and

non rigid animations. A real time 3D model reconstruction from a few images with adequate accuracy is still an outstanding challenge. This requires high computations and chances of losing information during reconstruction. Furthermore, lighting and illumination modeling makes the model more realistic and enables to synthesize against different lighting conditions. We propose light models as a possible extension of our approach and make the entire face model stable against varying poses, expressions and lighting conditions.

CHAPTER 6

Human Activity Recognition

In the previous three chapters, we have thoroughly studied analysis of human faces for interactive scenarios. In this chapter, we extend our work to visual analysis of action and activity recognition by using content information from action videos. Visual interpretation of human actions and activities has gained repute in current *Human Robot Interaction* (HRI) applications. Automatic analysis of human actions helps the robots to understand human behavior and these actions can be synthesized on the robots to perform joint activities with humans. Furthermore, it helps the robots to predict human intentions and manipulate everyday tasks to assist the humans. The future of the interactive technologies lies in the development of efficient algorithms which are capable to work in real time. For instance, *Microsoft kinect sensor* for full body actions [Mic], smart cameras with PC inside [Xim] which contain operating systems and processing capabilities, daily life interactive systems like vending machines commanded by human faces [Co.] and assistive robotics [Sys], where humans are assisted by robots to perform their routine tasks. The assistive scenarios are designed for elderlies and patients who need day long care without being interrupted. The assistive robots are developed for *care* purposes and are designed on the principles of artificial intelligence for automatic interaction with the humans. Classification of the human actions finds its applications in HRI, surveillance applications, assistive and medical robotics, gaming applications etc. In this chapter we study markerless action recognition from videos. We extend the study of content based video analysis in which a given action video is represented with a small number of feature descriptors. These descriptors correspond to the motion features in videos and are utilized to make vocabularies of different action components. Such approaches follow a bottom-up strategy where the contents are utilized to find a label of the given video. We study *bag of words* (BoW) in detail and propose that additional spatial features improve the classification rate on facial expressions recognition. We further study the effect of descriptor size on classification rate. A general overview of the different approaches used for action recognition is given in

Table 6.1.

The remaining part of the chapter is divided in four major sections. Section 6.1 gives brief introduction to action recognition techniques in computer vision by comparatively describing context and content based analysis of the videos. Section 6.2 gives overview of bag of word approach which describes interest point detection, vocabulary formation and feature extraction. In section 6.3 we present the results of the experimentation performed on facial expression recognition and activity recognition on TUM-kitchen dataset. Section 6.3.2 gives analysis of full body actions in different scenarios. Finally, section 6.4 concludes this chapter with some future extensions.

6.1 Introduction

Human activities can be decomposed in a sequence of different actions which are performed in a given context. They represent human intentions for joint interaction in HRI scenarios. Actions categorization depends upon the given applications. For instance, in assistive robotics, low level information is required in order to precisely and safely perform the joint tasks between humans and artificially intelligent agent. An example of the low level task is grasping objects which requires the precise position and localization of the object, hand tracking and gripping knowledge from the object geometry. Such tasks are usually interpreted from multi-sensor data along with visual data from cameras mounted either on the robots or in the environment. However, visual analysis is more intuitive, non-intrusive, less error prone because of controlled sensor noise and cost-effective. Everyday tasks manipulation for the robots treated in computer vision using two general approaches 1) top-down approach which is also called context-based image analysis and 2) bottom-up approach which is also called content based image analysis. Context information from a scene or video is utilized during the design phase. The context generally refers to the prior knowledge which contains different types of information including object geometry, object segmentation, silhouettes etc. In Chapter 4 and Chapter 5 we have studied context aware face image analysis using 2D and 3D models. In these approaches it is known that the given image contains a face by using a face detector. A model is designed which corresponds to local facial features and hence fully utilizes image context. From this top-level information, features are extracted and/or tracked throughout the activity and finally classified. Albeit the systematic design benefits, flexibility in adapting to different scenarios and detailed level information, these processes are relatively slow and require additional effort in object detection, segmentation, tracking and sometimes proper initializations [BEB08].

	Parametric Action Grammar	**Global Action Template**	**Local Bags of Features**
Parametric Body Model	Body Grammar	Body Posture	Bags of Postures
Global Image Model	Image Grammar	Image Template	Bag of Keyframes
Local Spatial Bag of Features	Features Grammar	Features Template	***Bag of Words***

TABLE 6.1 Overview of different methodologies used for action recognition as described by [WRB10].

On the other hand, content based image analysis exploits bottom-up strategies to evaluate the given visual data. As compared to aforementioned approaches, these approaches require no prior knowledge about the scene or object and hence do not necessarily require object detection, segmentation, tracking and initialization issues. Due to their design flexibility, ease in implementation and generalization, these approaches have the ability to work under noise and partial occlusions. In this chapter, we study human activity recognition using the similar approach and extract sparse spatiotemporal features from video data. These features are the descriptors of the local representations for the actions, not necessarily corresponding to specific body parts. A video or an image sequence is decomposed into smaller regions defined by the interest points. The sparseness of these interest points benefits in efficiency and robustness against occlusions and cluttered backgrounds. Such approaches exploit the content of an image or video by using a bottom-up strategy, which first detect the interest points and find a local descriptors to form a codebook for action classification [WRB10]. This approach is called *bag of words* (BoW) . Besides their several properties and representation strength in pattern classification application, there are several challenges in bag of words approach. Some of them include:

- **Vocabulary formation:** How to select a small and discriminative set of code words which form a vocabulary.
- **Feature combination:** How to combine different types of features. Additional spatial features improve the classification rate.
- **Loss of spatial information:** How to incorporate spatial information and position of the interest points. Interest points are chosen on the locations where the responses of the spatial and temporal filters are high. These locations do not necessarily corresponds to particular body parts but representative of their motions.
- **Multiple persons:** In case of multiple persons in the image, it is difficult to classify actions.

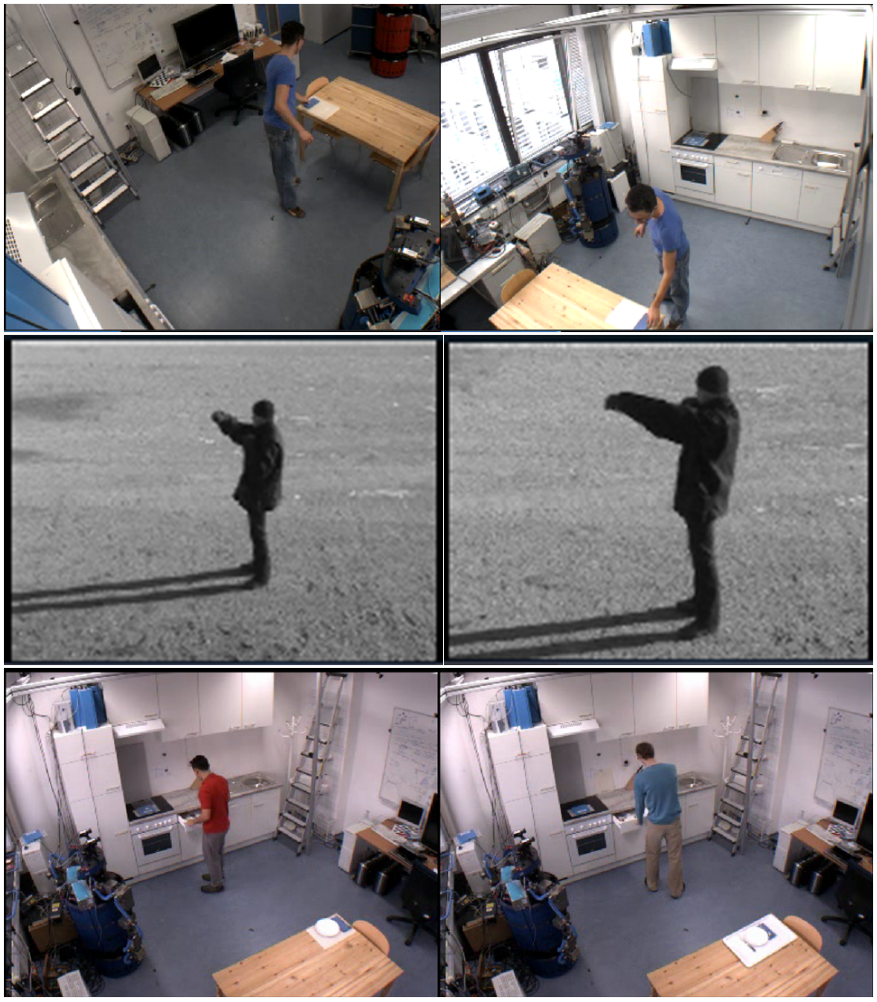


FIGURE 6.1 Different challenges in action recognition, (top row) Two different view points of the same action and also illumination changes, (middle row) different scaling, (bottom row) self occlusions in opening a drawer.

- **Rigid motions and background motions:** In case of spatiotemporal feature, it is hard to distinguish between foreground and background motions. Further rigid motion affects the facial expression classification results.
- **Monotonic and seamless actions:** To find robust interest points for monotonic and smooth motions instead of periodic motion components. For examples, table setting scenario from TUM-kitchen database [TBB09].

In addition to the aforementioned challenges, action recognitions have a few more challenges. Actions are performed differently by different persons and intra-personal variability may lead to misclassification. Further, different postures, size and illuminations may harm the

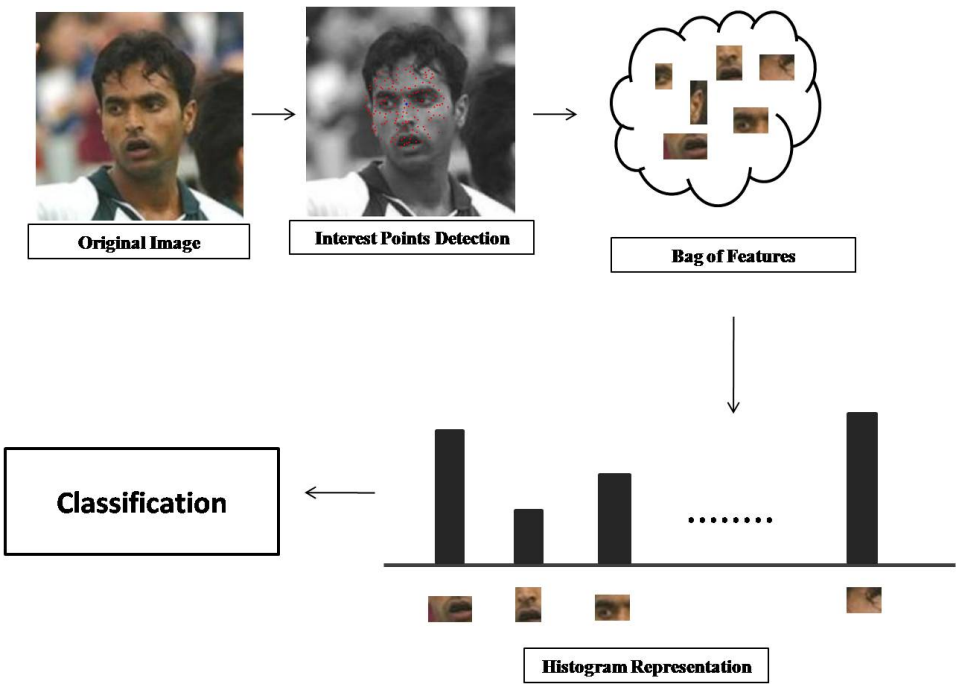


FIGURE 6.2 Conventional bag of visual words approach.

analysis of the activities. Actions usually contain temporal components and motion is performed either monotonically or periodically. This causes a change in background, occlusions and specially self-occlusions. These challenges are shown in Figure 6.1.

Bag of words approach was proposed by researchers [FFP05] as a useful tool toward automatic and unconstrained analysis of actions, object recognition and somehow to facial expressions. This method has been used by [Fei05][SRE+05]. We mainly analyze human actions in assistive environments by using TUM-kitchen dataset [TBB09].

6.2 Bag of Words (BoW)

Bag of Words (BoW) or *Bag of Features* (BoF) is one of the widely studied approaches in recent research because of its simplicity and representation power. This approach was proposed for document classification where a corpus of documents is represented by a compact set of vocabulary words. Each document represents *bag of words* which give rise to one or more topics regardless of their orders. This means that ``long life´´is same as ``life long´´. This shows that the content information is more useful than context in these scenarios. A relation between the topics and the documents is defined using *Latent Semantic Analysis (LSA)*, *Probabilistic Latent Semantic Analysis (pLSA)* or *Latent Dirichlet Allocation (LDA)*. From document

analysis, this idea is replicated to a set of images by finding meaningful regions or patches inside an image for vocabulary formation. A vocabulary is learned from the training data. One of the major advantages of BoW over other approaches is its application to the videos without prior knowledge and context. This generalization property and design flexibility make this approach suitable to apply on different image and videos classifications. In this chapter we study two applications; (1) actions recognition and (2) facial expressions classification. However from the experimental evaluations, we conclude that this approach is recommended for action recognition because of the higher global motions as compared to facial expressions where local motions are significant. Figure 6.2 shows generic picture of BoW. An image or a video is provided to the system and interest points are calculated independent of the objects in the image. These interest points usually represent corners, blobs or other meaningful regions which is discussed in section 6.2.1. We derive spatiotemporal features inspired from Dollar et al [DRCB05] for human action recognition. Further, optimization of the cuboid selection is discussed in detail. In addition, we add spatial features which enhance the performance of conventional approach. Figure 6.3 shows the process in detail with additional spatial features.

We use the approach proposed by Dollar et al. [DRCB05] which introduces the behavior analysis (facial expressions and action recognition) using sparse spatiotemporal features. A behavior is described as different types of the feature points instead of appearance, postures and occlusions. This makes the feature set robust against the most of the aforementioned challenges (as shown in Figure 6.1). Instead of finding counterparts of 2D interest points in 3D, an extension of *Harris* corner detector is applied to find the interest points in 3D. The third dimension is in temporal direction. Spatial and temporal filters are applied and non-maximal suppression is used from these filter responses to find the interest points at high filter response areas. The spatial filter is the *Gaussian* with σ ranging from $1.0 < \sigma < 3.0$. For temporal direction *Gabor* filters are used with $\tau = 1.5$. This spatiotemporal filter is designed in such a way that it can account for periodic motions. For instance, a boxing action consists of repeated arm motion in horizontal direction and hence interest points correspond to arm motion. A cuboid is extracted around each interest point. In order to find feature vector from the cuboids, gradient in all three directions are calculated which creates G_x, G_y and G_z blocks which are flattened to form a feature vector. A concatenation of the these blocks constitutes a feature vector. PCA is used to reduce the dimensions and this low dimensional data is clustered using k-means clustering to combine together the features of same types. Histograms of these cuboid types (which are called topics in BoW approach) are used as behavior descriptor. An SVM classifier is finally used for classification. For further details, refer to [DRCB05].

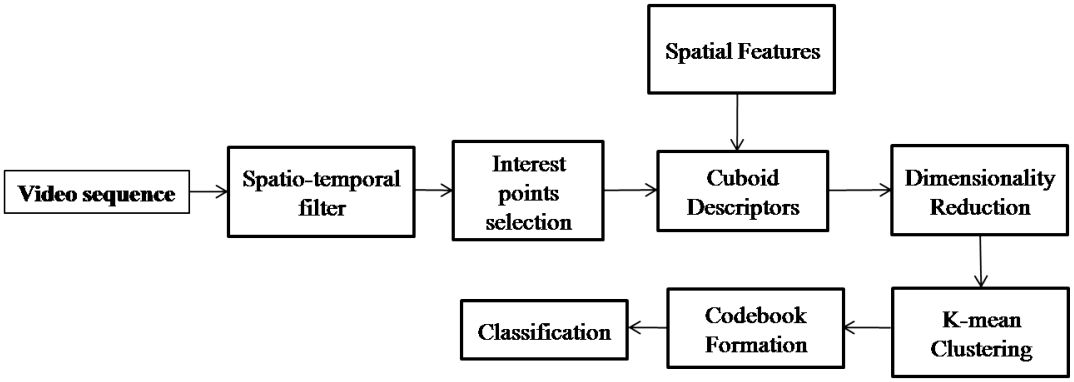


FIGURE 6.3 Sequential flow of BoW approach for action classification.

6.2.1 Interest point detection

In content based image analysis meaningful regions or patches are extracted which describe the whole image. This minimum set of patches provides compact representation of the given image. These regions are extracted from points in the image and these points are interchangeably called key points, salient points or interest points. Several approaches are used for this purpose, SIFT, *Harris* corners detector, *Kadir and Brady* detector [KB01b]. For details refer to [SMB00]. Grid based methods have also been used by the researchers however grid locations might not necessarily correspond to meaningful areas in the image. We apply two separable filters in spatial and temporal direction as proposed by [DRCB05]. A spatial filter is a *Gaussian* filter with a suitable σ value. This is applied in horizontal and vertical directions respectively with the same window size. A suggested value of σ is $1.0 < \sigma < 3.0$. A *Gaussian* filter response on a given image frame I_i in two direction is given by:

$$g(x, y : \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{6.1}$$

A temporal filter with a 1D *Gabor* quadrature pair is applied with τ ranging from $1.0 < \tau < 3.0$. The quadrature pair h_{ev} and h_{od} is applied using:

$$h_{ev}(t; \tau, w) = -cos(2\pi tw)e^{-\frac{t^2}{\tau^2}} \tag{6.2}$$

and odd filter using:

$$h_{od}(t; \tau, w) = -sin(2\pi tw)e^{-\frac{t^2}{\tau^2}} \tag{6.3}$$

Where $w = \frac{4}{\pi}$. The overall response function becomes $R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$. This makes R a function of two parameters σ and τ which represents spatial and temporal

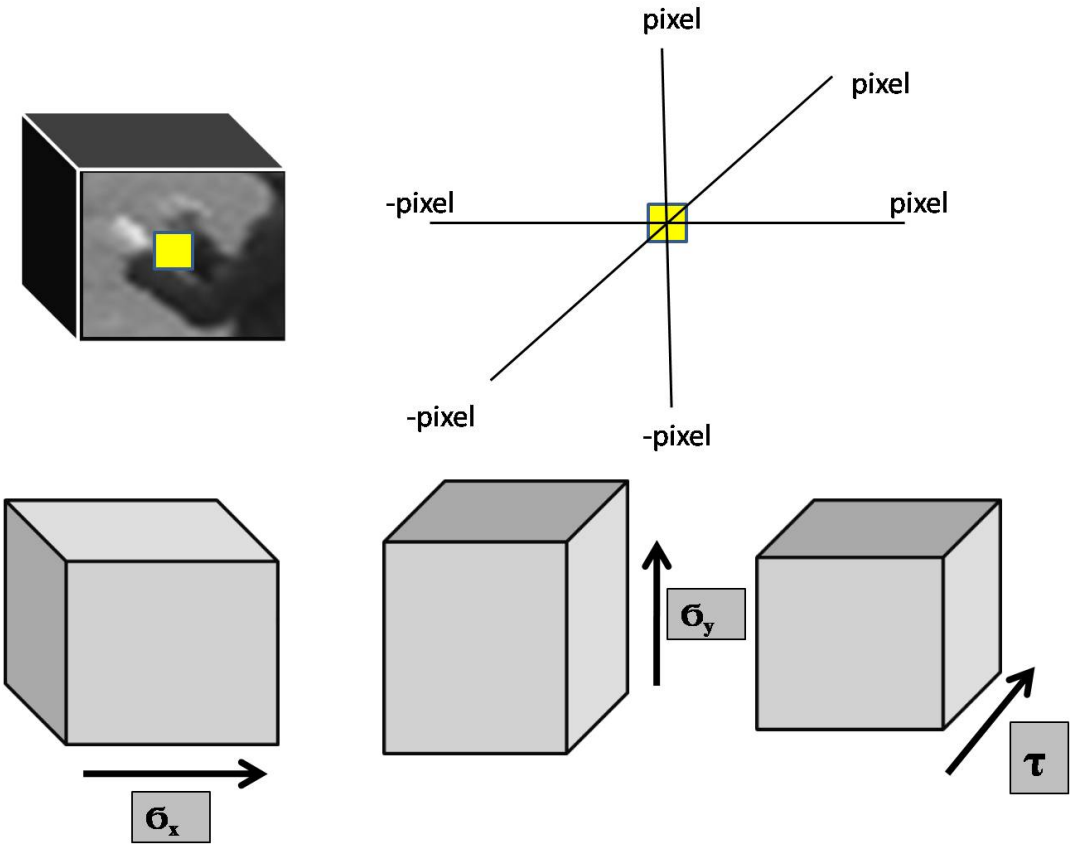


FIGURE 6.4 Cuboid orientation based on gradient values. Larger side of the cube reside along the higher gradient direction.

filters respectively. After smoothing in three dimensions, a cuboid is extracted with the size equal to $3\sigma \times 3\sigma \times 3\tau$. During experiments, we extract maximum 200 cuboids or 200 features. This is the maximum size of the cuboid chosen on the basis of error minimization.

We have experimented with a slight modification in the extraction of the cuboids by using a gradient constraint. At each interest point p_i, gradient value is extracted in all three directions. A cuboid is aligned with its largest side along the maximum gradient direction and smaller side in direction of smaller gradient. Figure 6.4 shows the orientation process in detail. For this we choose two σ values of spatial filter and hence the side of the cuboid is chosen as $3\sigma_1 \times 3\sigma_2 \times 3\tau$. Where σ_1 and σ_2 are two spatial filters in two spatial directions with higher value lying in the direction of higher gradient direction. The reason behind cuboid orientation is to optimally select the best feature descriptor. For example in case of a boxing sequence, hand motion is taking place mainly in horizontal direction. So the cuboids around the interest points should be extracted in such a way that the maximum motion is captured by each cuboid.

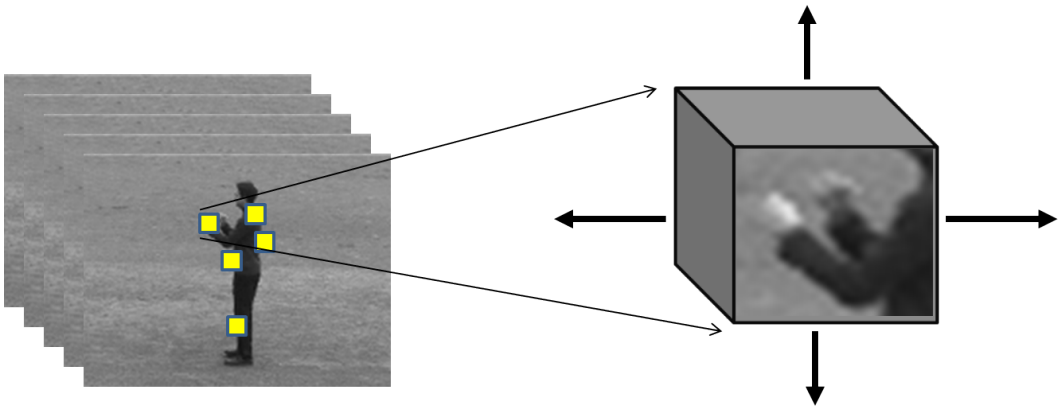


FIGURE 6.5 Example of a boxing action. Most of the action components occur along horizontal axis. A very less or slight motion occurs along vertical axis.

In this way largest side of the cuboid should reside along the horizontal axis. Figure 6.5 shows this evidence in detail. We obtain a slight improvement on some codebook sizes for action recognition. Furthermore, by using additional spatial features (DCT coefficients and 3D SIFT features), we obtain better results as compared to conventional BoW approach [DRCB05] for facial expression recognition. We have also obtained better results on TUM-kitchen dataset as compared to modeling full body for action recognition [TBB09].

6.2.2 Feature Descriptor and Vocabulary Formation

Once defined, the size of the cuboids is not changed throughout the experimentation. Once a cuboid is extracted, it is aligned as a row vector and used for PCA based dimensionality reduction. The reduced data is used for *k-means* clustering to form a vocabulary of the cuboids prototypes. We used different codebook sizes during experimentation and analysis. A holistic approach is used for the action recognition. A video sequence consists of a single action performed from start to end and repeated several times in case of kth-database videos. We initialize by applying interest points detector to find points of interest inside a video. These interest points correspond to peak response of the spatiotemporal filter. We choose maximum 200 interest points from all video sequences. If a video shows less than 200 points then all those points are considered. Higher number of points are truncated to 200. Locations of these points are saved along with the respective video. A *Gaussian* filter with $\sigma = 1.5$ is applied as spatial filter and $\tau = 2.5$ is applied as temporal filter on these interest points. After spatiotemporal filtering, a small cuboid is extracted from these locations. The size of these cuboids is chosen on the basis of spatiotemporal filters. We extract a cube of size

$ceil\ \{3\sigma\ +\ 1\}\ \times ceil\ \{3\sigma\ +\ 1\} \times ceil\ \{3 * \tau\ +\ 1\}$.

In case of oriented cuboids, we choose *Gaussian* filter with two different variations: $\sigma_1 = 1.0$ and $\sigma_2 = 1.5$ whereas temporal filter is not changed. Once the maximum gradient direction is found, we orient the cube to that direction as explained in section 6.2.1. The cuboid size is $ceil\ \{3\sigma_1 + 1\}\ \times ceil\ \{3\sigma_2 + 1\} \times ceil\ \{3 * \tau + 1\}$. All cubes are smoothed with two different variations of *Gaussian* filters. These two variations include $\sigma = 1.5$ and $\sigma = 2.5$. All pixel values are extracted from this cube and concatenated to form a raw feature vector. We store all these vectors as a matrix of size $N \times M$, where *N* is the number of interest points in whole database and *M* is the size of the raw feature vector. Finally a PCA is executed over the whole matrix to reduce the dimensions. On this reduced dataset, *k-means* clustering is performed. We chose *k* with different values to study the effect of different codebook sizes. We obtain k-words in the vocabulary. The histograms are formed as a final feature set and classification. feature extraction process in detail.

6.3 Experimental Evaluations

For experimentation purpose, we choose two different databases. Action and activity recognition is conducted mainly on TUM-kitchen database and facial expressions classification is performed on University of California database. We use SVM for classification with histogram intersection kernel and leave one out cross validation approach. In this way number of folds for cross validation is equal to the number of persons in the database. Codebooks are formed using video words which are extracted after k-means clustering. For facial expressions recognition, training and testing is performed by keeping one person in training and other in testing for six different facial expressions. This database is recorded under controlled conditions as compared to CKFE database [KCT00] and MMI database [MSV+09]. Facial expressions videos are captured for six standard facial expressions anger, disgust, fear, laugh, sad and surprise. Each video starts with a neutral face and ends again on neutral expression but between these points facial expressions rise to their peak. The faces are captured against a dark background and hence there is no background variation or clutter to deteriorate the classification results. Furthermore, there is no rigid motion in the person's head and it remains almost stationary throughout the expression generation. Figure 6.6 shows some frames from the videos of this database with detected interest points and cuboids.

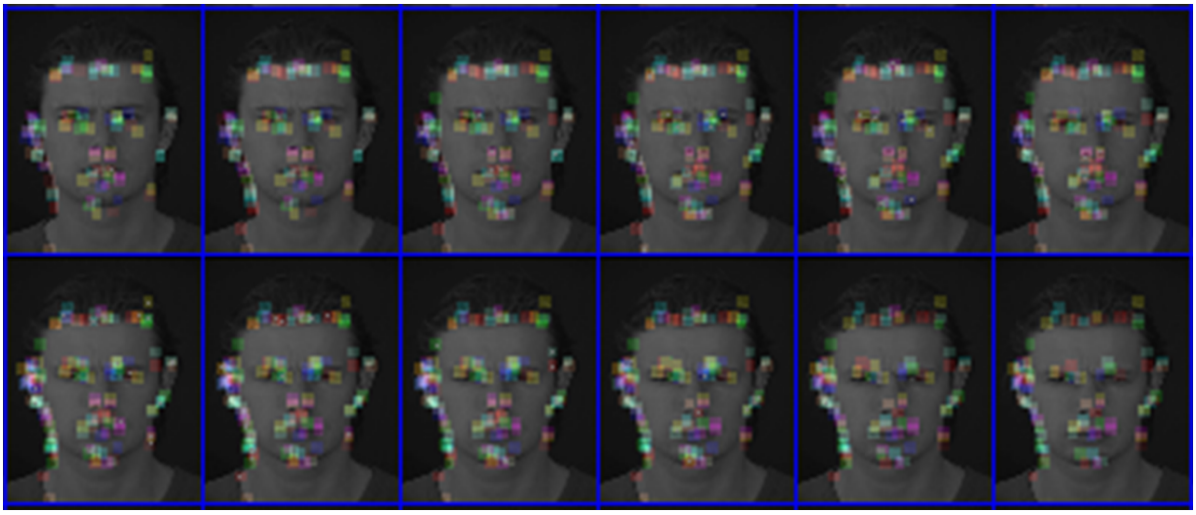


FIGURE 6.6 Examples of interest point detection on a facial expressions sequence with cuboid descriptor.

6.3.1 Facial Expressions Recognition

We experiment facial expression recognition on the dataset used by Dollar et al. [DRCB05] at University of California. This database represents six universal facial expressions which are, anger, disgust, fear, laugh, sadness and surprise. The database is captured with two persons against a dark background under controlled conditions. The subjects in the database do not perform any global motions. From the experimental evaluations, it can be concluded that under these controlled conditions, BoW can be successfully applied. However, this approach does not perform satisfactorily on MMI database [MSV+09] and CKFE database [KCT00]. For experimental evaluations, we consider one person in training and other in testing and compared our results with conventional BoW approach. We obtain an improvements in the results by adding additional spatial features. The cuboid orientation methodology is given in detail in section 6.2.1 whereas DCT feature extraction process is same as given in section 3.3.2. Besides the regular spatiotemporal features described in section 6.2.2, we add first five DCT coefficients from each cuboid layer and combine a feature vector after feature normalization. Features are normalized such that they have unit norm after normalization. We choose SVM as a classifier with histogram intersection kernel from LibSVM [CL01]. Figure 6.7 shows the results of facial expression recognition using DCT coefficients and temporal parameters. Figure 6.8 shows comparison of approach used by Dollar et al [DRCB05].

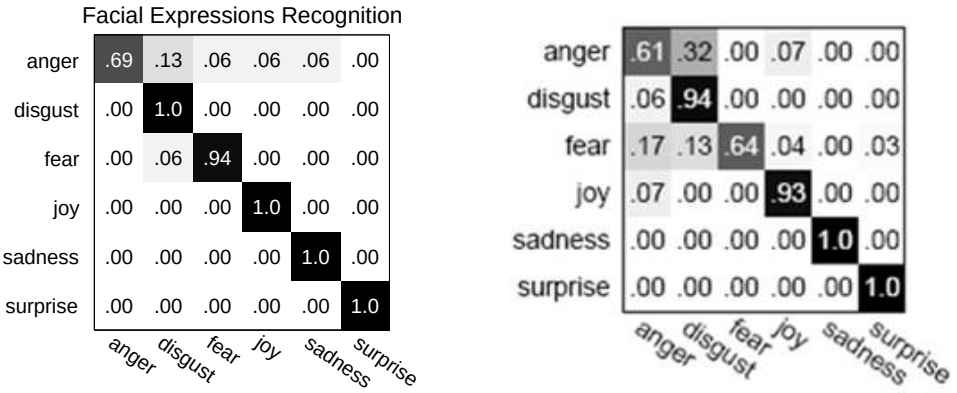


FIGURE 6.7 Facial expressions recognition results using our approach (left) and the approach used by Dollar et al. [DRCB05]

6.3.2 Full Body Action Recognition

In order to study bag of words approach for action recognition, we use an assistive environment where different persons are performing activities under an observed scenario. We use TUM-kitchen database [TBB09] which consists of a table setting activity performed by different persons and performance of the same activity by the humans acting as a robot. The database provides sufficient variations to study action classification which are useful for manipulation of the high level tasks. These high level tasks include pick and place objects, opening door etc. The videos are captured from four ceiling cameras at four corners of a cubic room. This database is smaller in size as compare to other action recognition databases like kth-database and Weizmann-database however self occlusions and action complexity are high in this database. The activities are performed by different actors in their natural way and these activities are also performed again by the same actors in such a way that these activities are now synthesized by the robots. A single person enters the observed environment and starts placing different objects on the table while repeating few activities. These tasks are performed by different actors in different ways. Finally, after setting the table person leaves the environment. These activities are observed with all four cameras mounted on the corners of the ceiling. Each video represents single activity which further consists of several actions. For details refer to [TBB09].

Each video represents a single person in observation. Furthermore, the background is cluttered but stationary. We divide each video manually in eight different tasks: entering, exiting, opening door, closing door, picking object, placing object, opening drawer and closing drawer. Figure 6.10 shows example of these actions performed inside kitchen environment with *cam-*

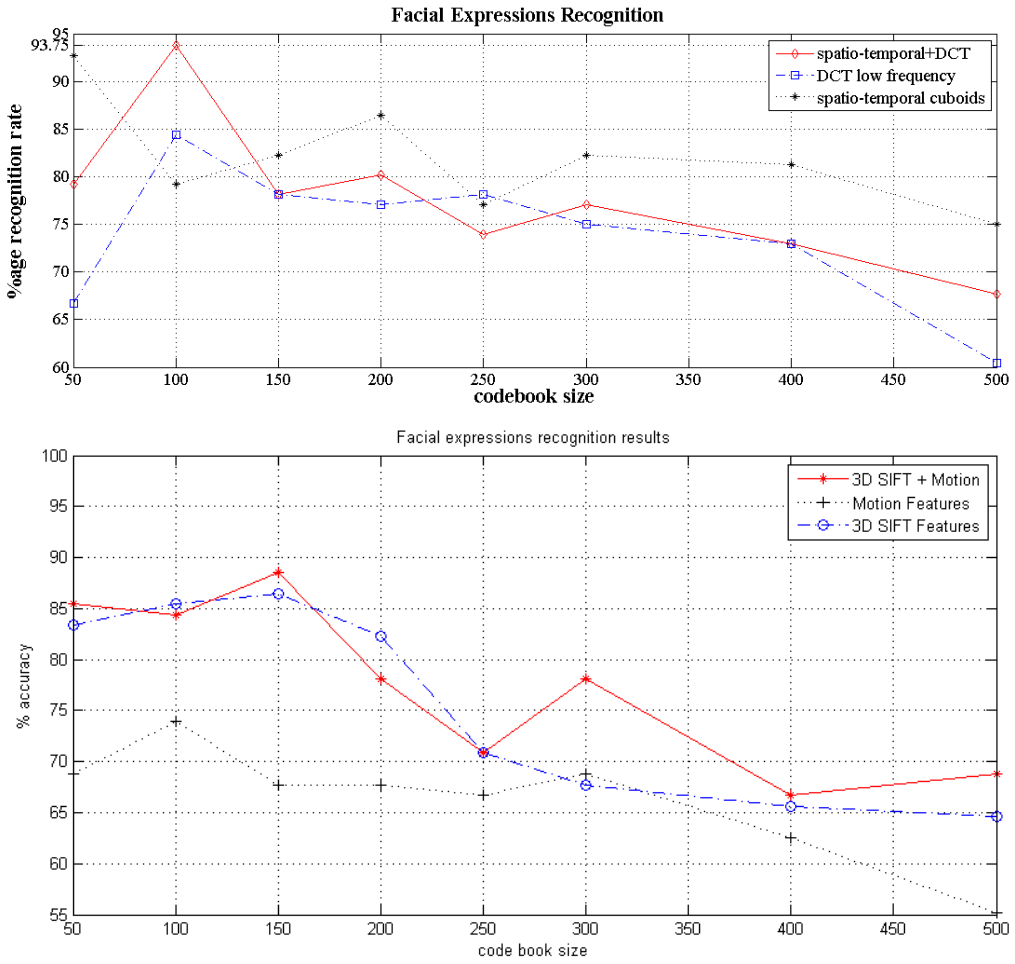


FIGURE 6.8 Facial expressions recognition comparison using cuboids, DCT coefficients and a combination of both (above). Facial expressions recognition comparison using cuboids, 3D SIFT and a combination of both (below).

era 3. The results on this database using BoW outperforms the results given by the Tenorth et al. [TBB09] using model based approach.

The results show the capability of the BoW to work in the presence of self occlusions and cluttered backgrounds. The action classification results from *camera 0* and *camera 1* show better performance since most of the activities are performed in the field of view of these cameras. These activities are occluded in the other two camera views and hence classification rate decreases. We again use SVM with histogram intersection kernel from LibSVM [CL01] for classification and use leave one out cross validation approach.

In addition to conventional BoW for actions recognition on table setting activity data, we have studied the effects of the different cuboid sizes on the action recognition on a subset

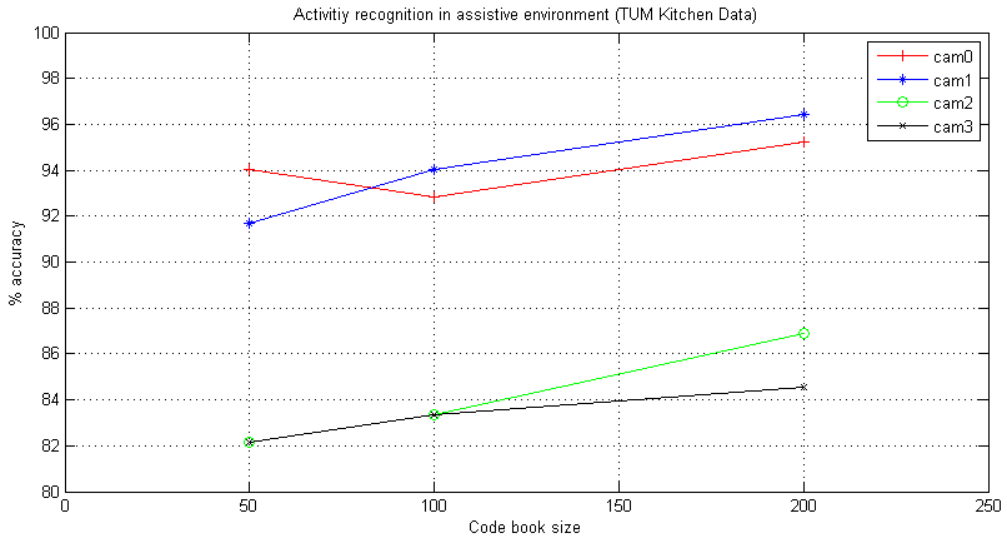


FIGURE 6.9 Action recognition on kitchen database from four different camera views. It can be seen that the person independent approach works even in the presence of self-occlusions.

of kth-database . We randomly choose 10 subjects and applied two different algorithms to get the results. Figure 6.11 shows the comparison of conventional BoW to the varying cuboids methodology. From this figure we can see that for different codebook sizes, we obtain slightly better results.

6.4 Conclusions and Future Work

This chapter addresses a state of the art approach toward full body action recognition and facial expressions classification. This approach is called *bag of words* (BoW) where the words are extracted from motion features in a videos. The approach followed in this chapter is adapted from Dollar et al. [DRCB05]. We experiment with: 1) additional spatial features and 2)oriented cuboids to study the areas of improvements. We conclude with:

- One of the major benefits of BoW is that it does not require any context information like person detection, segmentation and tracking. Further it is markerless approach and hence can be applied for body language analysis for HRI applications.
- This approach is useful for higher level actions which are generally present in the action recognition databases. These actions are, for example walking, boxing, jumping etc., in controlled conditions. Besides these tasks, this approach can also help to manipulate low level associated tasks.

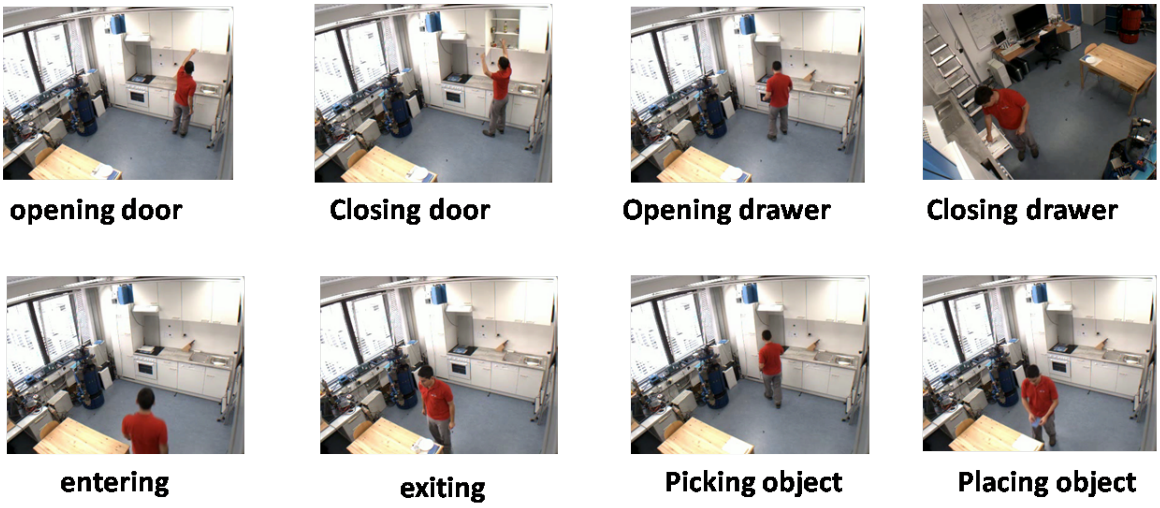


FIGURE 6.10 Eight classes of action in assistive kitchen environment, where a person is setting a table.

- Experimental results show that the additional spatial features or a set of multiple features enhances the classification performance.
- A cuboid descriptor can be further enhanced by increasing the size of the raw feature vectors. This is achieved by using cuboid orientation and capturing the direction of higher gradient.
- Bag of video words is suitable for the situations where the rigid motion of the objects is limited. Hence for facial expressions if head motion is higher then most of the interest points correspond to head boundary rather than local facial features. This effect causes more confusion in different facial expressions.

The real time performance of BoW approach still requires attention from the research community. Further the classification rate is affected with higher background motions and cluttered environments.

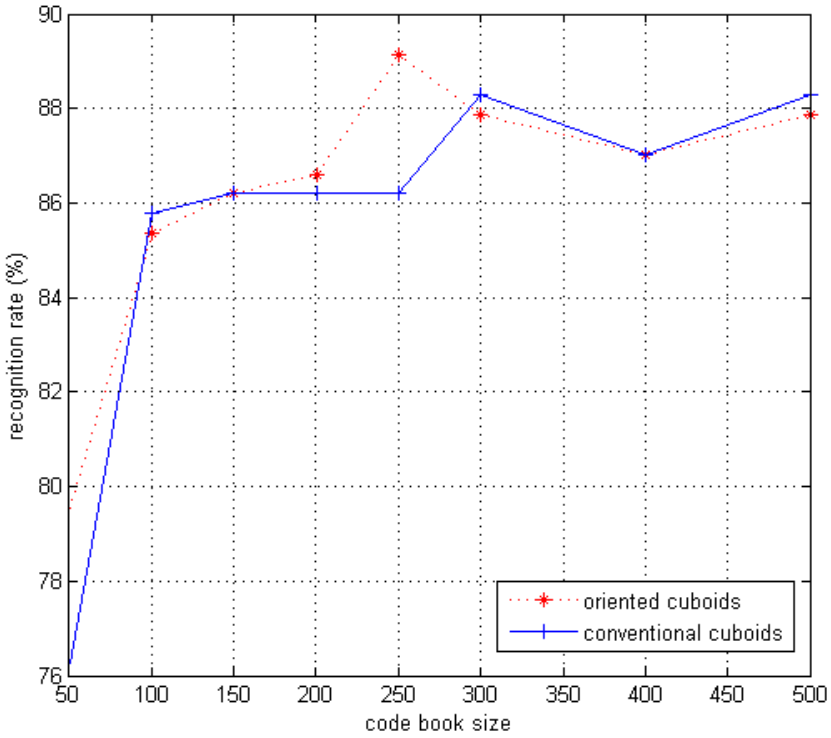


FIGURE 6.11 Action recognition from kth-database database using oriented cuboids.

CHAPTER 7

Conclusion and Future Directions

In this thesis we have studied visual analysis of body language for *Human Robot Interaction* (HRI) applications by explaining its two different aspects, i.e. 1) face image analysis and 2) full body actions and activity recognition. The main goal of this research work is to extract a robust feature set for visual interpretation of body language in these scenarios. This feature set is capable to work in real world environments. In analysis of body language, we extract facial identity, facial expressions, gender, estimation of the age and classification of activities performed by the humans. Section 7.1 summarizes the approaches followed in this thesis. In section 7.2 we present some conclusions drawn from this work and finally section 7.3 gives possible future extensions of this work.

7.1 Summary

In this thesis, we have mainly contributed toward robust feature extraction techniques to study human behavior during interaction with the robots. For face image analysis, 2D and 3D face models are studied in detail. A comparative study of these models shows that 3D realistic model outperforms 2D appearance model. We propose a feature set called *Spatiotemporal Multiple Feature* (STMF) . This feature set is configured from structural, textural and temporal components of the face image sequences. This feature set is extracted from both 2D and 3D models. The general configuration of the feature set remains the same for two types of models but improvements are made in structural and textural components of the feature set for its 3D counterpart. The richness and representative strength of this feature set is tested by using it for face recognition, facial expressions classification, gender classification and age estimation. This image based feature set is capable to classify different attributes of a face and stable against varying poses and facial expressions. We propose this set for the socially inspired robots where robots are the part of the environment and capable to extract multiple

information from a given face by using a single feature set. In order to assure the robustness of the proposed feature set, we have tested it in real world scenarios. We evaluate the performance of this feature set against varying poses and facial expressions in a face recognition system. The performance of the recognition system improves by using a 3D model in the presence of varying poses and facial expressions. Moreover, the effects of perspective distortions are ignored by the researchers on the face images for face recognition systems because the distance of the camera to the face is larger than the face size. We have studied the effects of perspective distortion on face regions and report a better performance in face recognition and age estimation. This is achieved by using a 3D wireframe model. The texture is extracted from face images by using 3D model and stored in a texture map after rectifications. Each triangular patch is given equal weight in this texture map. In addition to model based approaches, we have also studied conventional approaches for face recognition and feature sparseness. These approaches are error prone in the presence of lighting variations and head rotations, therefore we have used a general method for the preprocessing of the images for feature extraction. This stabilizes the system for real time performance. Besides conventional approaches, we have also studied sparse features for face recognition. They improve the classification rate as compared to holistic approaches on standard databases. We extract these sparse features from interest points and grid points. Finally we have studied *bag of words* (BoW) approach for full body action recognition in an assistive environment. This approach is stable against occlusions and view invariant. A major benefit of this approach is that it does not require prior knowledge about the object and works without object segmentation, tracking and initialization. We have studied this approach for facial expressions recognition and full body action recognition. The performance of a facial expression system is improved by using additional spatial information from *Discrete Cosine Transform* (DCT) coefficients and *3D SIFT* features. This approach also outperforms the model based interpretation of full body actions in a table setting scenario.

7.2 Conclusions

We have studied different feature extraction approaches in this thesis for human robot interactive applications. It is quite useful for the robots to understand human behavior and manipulate everyday human activities to facilitate elderlies and the patients who need a special attention. For this purpose, we have studied different features which have a close relation to human visual capabilities and robust to work in our daily life activities. In assistive scenarios, cognitive capabilities in the robots enable them to analyze everyday human tasks and manipulate these tasks when performing joint activities with the robots. We have studied context aware

approaches for human face image analysis and a content based approach for full body actions. We conclude with the following observations and recommendations:

- A model based spatiotemporal feature set provides sufficient strength to classify human faces, facial expressions, gender classification and estimation of the age. This feature set consists of structural, textural and temporal variations of the faces and robust against varying poses and facial expressions for a face recognition system.
- The quality of the given feature set is studied with two different face models in 2D and 3D space. The general configuration of the feature set remains same in both cases however features extracted from a 3D model have more representation power as compared to 2D models. A 3D model outperforms comparative 2D model because of detailed texture and structural information.
- In conventional AAM, texture is warped from a given face to a target face using piecewise affine transformation. This transformation preserves collinearity and do not rectify image texture against perspective distortions. A well realized texture map can improve the texture features which give better performance than conventional 2D appearance models.
- The face recognition system developed using this approach is stable against varying poses and facial expressions. Such systems are useful for any HRI application because human faces are always seen in action conveying different expressions with different facial poses.
- We further concluded that sparse feature representation of the face images outperforms holistic representation. Features are extracted from the given image using interest points which correspond to some meaningful area in the image. We used *Harris* corner detector for interest point detection from entropy coded images.
- For action and activity recognition, we study *bag of words* (BoW) approach with a special focus on feature descriptor. By using additional spatial features and changing the size of the cuboid have the capability to give improved results.

7.3 Future Work and Extensions

This work is designed for assistive robotics where robots assist humans especially elderly and patients to perform their daily life activities. This research work can also be applied to other

FIGURE 7.1 The proposed face model is testes on frontal, half-profile, pitched up and pitched down faces with limited lighting variations. Our future goal is to fit the model to full profile and with varying lighting conditions as shown in this Figure.

daily life interactive systems. So far we have studied joint interaction using a single modality in an unconstrained environment. However, in such applications data acquired from multiple sensors is more informative and can further improve the performance of the system. Although the use of extra sensors increase the cost of the analysis but more informative. Further the face model is studied using a surface mesh which realizes the texture in a detailed way. A photorealistic model with light model can deal with the lighting variations and make the system capable to work under varying poses, facial expressions and lighting variations. Lighting models like *Phong reflection models* [Bus] consider the effects of three different lightings on the face region. These three components are ambient, diffused and specular lights. A future extension of our 3D model is to consider light model and synthesize more realistic views in different environments. Furthermore, the proposed feature set has been successfully used for face recognition, facial expressions, gender classification and age estimation. It is also recommended to use these features for ethnicity classification. Similarly, the proposed features are also recommended to be used in face pair matching, facial authentications and watch list check applications of *Biometrics* area. The 3D face model used in this thesis is generated from a single image. The model is projected to the 2D image and fitted to the face image. This model is reprojected to the 3D space and third dimension is approximated. In future we plan to use multiple views of a single face to generate its model. The depth information from range data also provides detailed realistic model.

For full body action recognition using sparse spatiotemporal features, we study that the performance of the recognition system can be improved by using detailed spatial features and optimal choice of the feature descriptor. In general, activities take place in the form of two or more seamless actions performed in a given context. In such situations, it is hard to identify that where does a particular action starts and where does it finish? In future, we also plan to apply this approach to seamless and longer activities classification. The approaches proposed in this thesis are partially implemented on a humanoid robot. In future, we plan to release this work for research community to directly apply it to robotic systems for standardization of different tasks performed by robots.

APPENDIX 8

Screenshots

FIGURE 8.1 Examples of interest points from entropy coded images from LFW database .

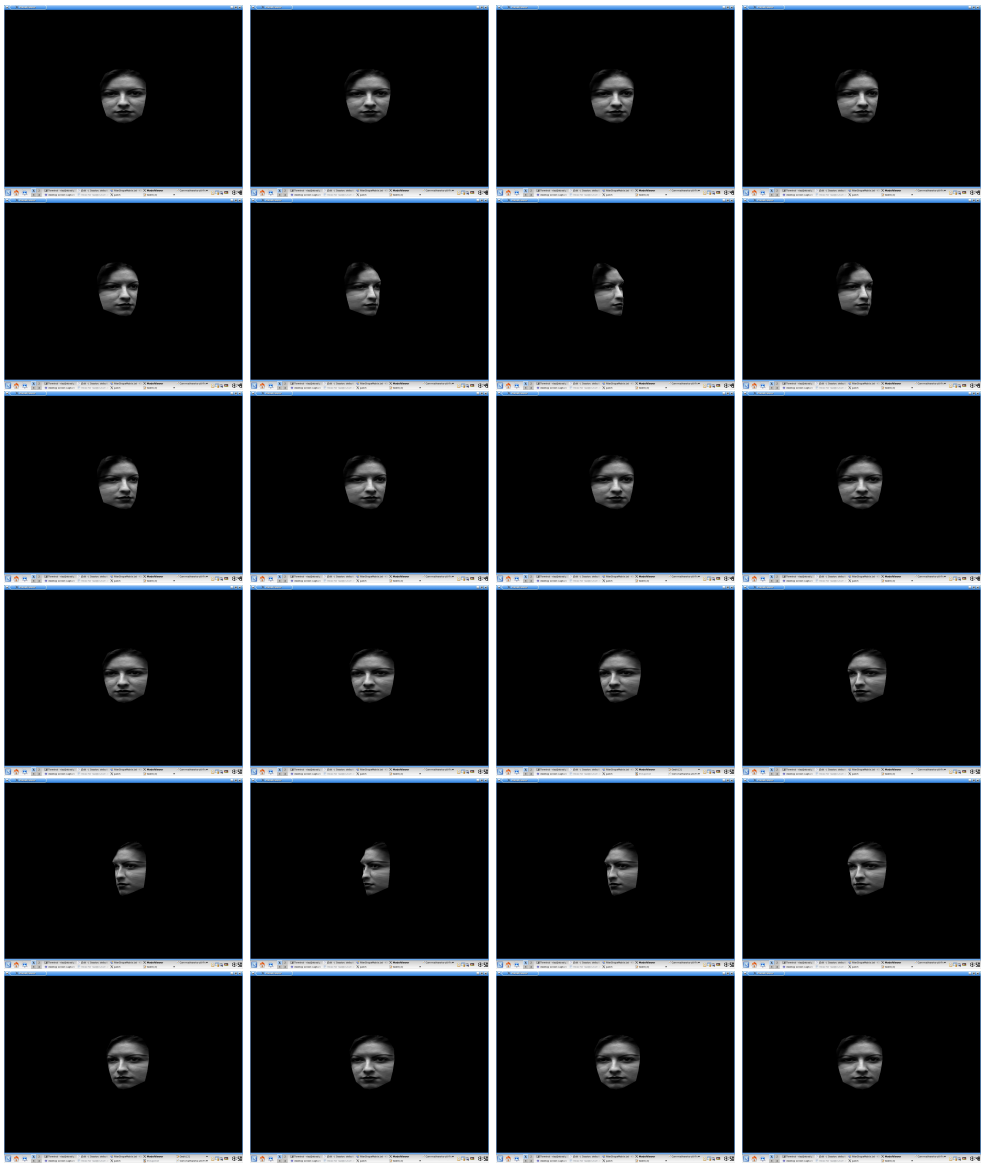


FIGURE 8.2 An example of the 3D realistic face model generated from a single image. The model is showing global motions.

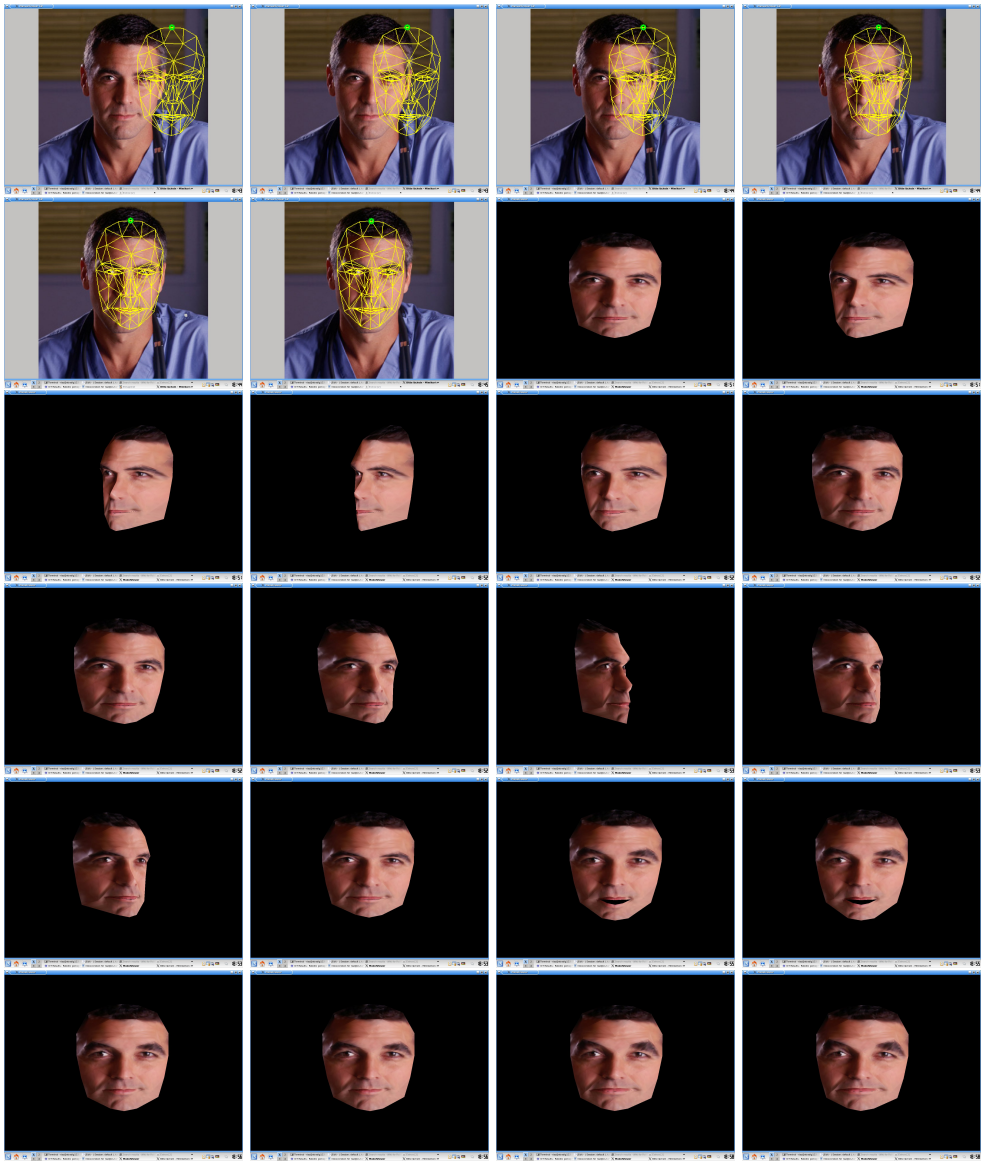


FIGURE 8.3 An example of the 3D realistic face model generated from a single image. The model is showing global motions and local deformations caused by FACS animation vectors.

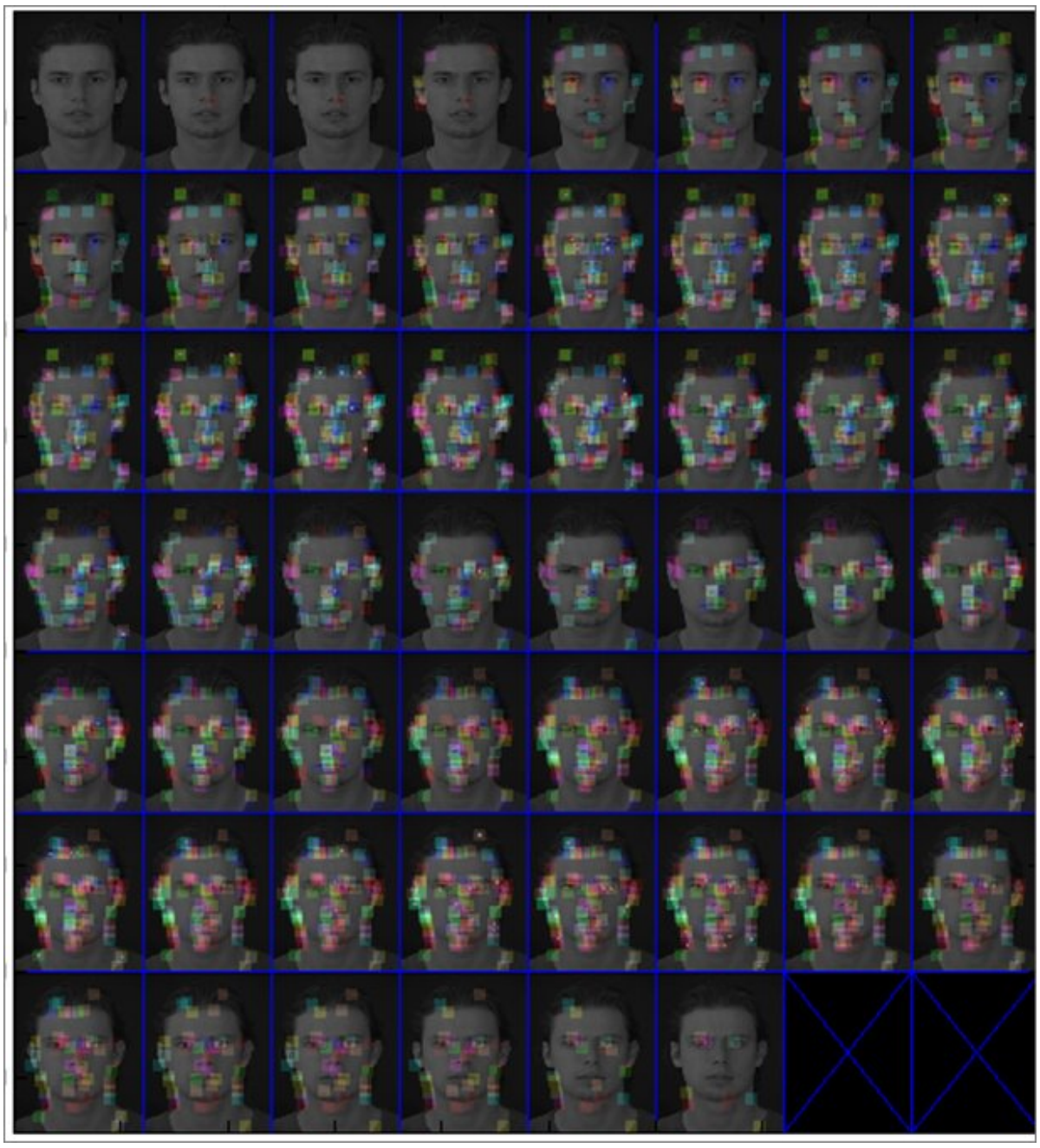


FIGURE 8.4 Examples of interest points for facial expressions recognition using BoW .

Bibliography

[ABV09] B. Amberg, A. Blake, and T. Vetter. On compositional image alignment, with an application to active appearance models, 2009.

[Ahl00] Jörgen Ahlberg. Candide - a parameterized face. http://www.icg.isy.liu.se/candide/, 2000.

[Ahl01] Jörgen Ahlberg. An experiment on 3d face model adaptation using the active appearance algorithm. *Image Coding Group, Deptt of Electric Engineering, Linköping University*, 2001.

[ALC+09] Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F. Cohn, Tsuhan Chen, Zara Ambadar, Kenneth M. Prkachin, and Patricia E. Solomon. The painful face - pain expression recognition using active appearance models. *Image Vision Comput.*, 27:1788–1796, November 2009.

[AMH+06a] Timo Ahonen, Student Member, Abdenour Hadid, Matti Pietikäinen, and Senior Member. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:2037–2041, 2006.

[AMH+06b] Timo Ahonen, Student Member, Abdenour Hadid, Matti Pietikäinen, and Senior Member. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:2037–2041, 2006.

[AMM+08] Alberto Albiol, David Monzo, Antoine Martin, Jorge Sastre, and Antonio Albiol. Face recognition using hog-ebgm. *Pattern Recognition Letters*, 29(10):1537–1543, July 2008.

[ASWG09] Akshay Asthana, Jason Saragih, Micheal Wagner, and Roland Goecke. Evaluating aam fitting methods for facial expression recognition. *Affective Computing and Intelligent Interaction*, 1(1):598–605, September 2009.

[BBK03] Alexander M. Bronstein, Michael M. Bronstein, and Ron Kimmel. Expression-Invariant 3D Face Recognition, 2003.

[BDF+03] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, March 2003.

[BEB08] Jan Bandouch, Florian Engstler, and Michael Beetz. Evaluation of hierarchical sampling strategies in 3d human pose estimation. In *Proceedings of the 19th British Machine Vision Conference (BMVC)*, 2008.

[Ber09] Bertillon system. *Encyclopedia Britannica*, pages 33–36, November 2009.

[BHK97] Peter N. Belhumeur, João P. Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, 1997.

[BI98] Andrew Blake and Michael Isard. *Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion*. Springer, 1998.

[BLGT06] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli. On the use of sift features for face authentication. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on*, pages 35 – 35, 2006.

[BMS02] Marian Stewart Bartlett, Javier R. Movellan, and Terrence J. Sejnowski. Face recognition by independent component analysis. *IEEE Transactions on Neural Networks*, pages 1450–1464, 2002.

[Bou00] Jean-Yves Bouguet. Pyramidal implementation of the lucas kanade feature tracker, 2000.

[Bou05a] R. Bouckaert. Bayesian network classifiers in weka. Technical report, Technical Report, Department of Computer Science, Waikato University, Hamilton, NZ, 2005.

[Bou05b] Remco R. Bouckaert. Bayesian network classifiers in weka. Technical report, Department of Computer Science, Waikato University, Hamilton, NZ, 2005.

[Bra01] Matthew Brand. Morphable 3d models from video. In *CVPR (2)* [Bra01], pages 456–463.

[Bru09] Roberto Brunelli. *Template Matching Techniques in Computer Vision: Theory and Practice*. Wiley Publishing, 2009.

[BT05] J.L. Barron and N.A. Thacker. Tutorial: Computing 2d and 3d optical flow. Technical Report Tina Memo No. 2004-012, 2005.

[BTG06] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *In ECCV*, pages 404–417, 2006.

[Bus] S.R. Buss. *3D Computer Graphics: A Mathematical Introduction with OpenGL*. Cambridge University Press. 2003.

[BV03a] Curzio Basso and Thomas Vetter. Regularized 3d morphable models. In *Proceedings of Higher-Level Knowledge in 3D Modeling and Motion Analysis*, pages 3–11, 2003.

[BV03b] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003.

[CET98a] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active Appearance Models. *Proceedings of the European Conference on Computer Vision*, 2:484–498, 1998.

[CET98b] TF Cootes, G Edwards, and CJ Taylor. Active appearance models. *Proceedings of European Conference on Computer Vision*, 2:484–498, 1998.

[CET01] TF Cootes, G Edwards, and CJ Taylor. Active appearance models. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 23:681–685, 2001.

[CH92] Gregory F. Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Mach. Learn.*, 9:309–347, October 1992.

[CHC+94] TF. Cootes, A. Hill, C.J.Taylor, J.Haslam, and Manchester M Pt. The use of active shape models for locating structures in medical images, 1994.

[CL01] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[Co.] JR East Water Business Co. Vending machines for drink selection. http://www.jreast.co.jp/.

[CSG+03] Ira Cohen, N Sebe, A Garg, Chen Lawrence, and Thomas Huanga. Facial expression recognition from video sequences: temporal and static modeling. In *Computer Vision and Image Understanding*, pages 160–187. Elsevier Inc., 2003.

[CT01] T. F. Cootes and C.J. Taylor. On representing edge structure for model matching. In *In CVPR*, pages 1114–1119, 2001.

[CTCG95] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models: their training and application. *Comput. Vis. Image Underst.*, 61:38–59, January 1995.

[CYTS10] Z.M. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *Computer Vision and Pattern Recognition*, pages 2707–2714, 2010.

[Dar73] Charles Darwin. *The expression of the emotions in man and animals*. D. Appleton, New York, 1873.

[DRCB05] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 14th International Conference on Computer Communications and Networks*, pages 65–72, Washington, DC, USA, 2005. IEEE Computer Society.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *In CVPR*, pages 886–893, 2005.

[ea07] Beetz M. et. al. The assistive kitchen — a demonstration scenario for cognitive technical systems. *Proceedings of the 4th COE Workshop on Human Adaptive Mechatronics (HAM)*, 2007.

[ECT98] Gareth J. Edwards, Timothy F. Cootes, and Christopher J. Taylor. Face recognition using active appearance models. In *Proceedings of the 5th European Conference on Computer Vision-Volume II - Volume II*, ECCV '98, pages 581–595, London, UK, 1998. Springer-Verlag.

[EF78] Paul Ekman and Wallace Friesen. The facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press*, 1978.

[EFH02] Paul Ekman, Wallace V. Friesen, and Joseph C. Hager. Facial action coding system (facs). *Salt Lake City (USA)*, 2002.

[Eke09] Hazim Kemal Ekenel. *A Robust Face Recognition Algorithm for Real-World Applications*. PhD thesis, Department of Computer Science, University of Karsruhe (TH), February 2009.

[ELTC96] G. J. Edwards, A. Lanitis, C.J. Taylor, and T. F. Cootes. Statistical models of face images - improving specificity. In *In British Machine Vision Conference*, pages 765–774, 1996.

[ESTS09] Hazim Kemal Ekenel, Lorant Szasz Toth, and Rainer Stiefelhagen. Open-set face recognition-based visitor interface system. In *Proceedings of the 7th International Conference on Computer Vision Systems: Computer Vision Systems*, ICVS 09, pages 43–52, Berlin, Heidelberg, 2009. Springer-Verlag.

[ETC98] G. J. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In *Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, FG '98, pages 300–, Washington, DC, USA, 1998. IEEE Computer Society.

[Fac] FaceGen. facegen. *Face Gen*, 2010.

[FCGH08] Yun Fu, Liangliang Cao, Guodong Guo, and Thomas S. Huang. Multiple feature fusion by subspace learning. In *CIVR*, pages 127–134, 2008.

[Fei05] *A Bayesian hierarchical model for learning natural scene categories*, volume 2, Washington, DC, USA, June 2005. IEEE Computer Society.

[FFP05] Li. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. *CVPR*, pages 524–531, 2005.

[fgn] Fg-net aging database. http://www.fgnet.rsunit.com/.

[GBK01] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.

[GBS+07] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.

[GRCC04] H. Gupta, A.K. Roy Chowdhury, and R. Chellapa. Contour-based 3d face modeling from a monocular video. pages xx–yy, 2004.

[Hec86] Paul Heckbert. Survey of texture mapping. In *IEEE Computer Graphics and Applications*, pages 56–67, November 1986.

[Hec95] D. Heckerman. A Tutorial on Learning with Bayesian Networks. Technical report, Microsoft Research, Redmond, Washington, 1995.

[HHB03] Jennifer Huang, Bernd Heisele, and Volker Blanz. Component-based face recognition with 3d morphable models. *International conference on audio and video-based person authetication (AVBPA)*, 3(9):27–34, 2003.

[HL01] Ziad M. Hafed and Martin D. Levine. Face recognition using the discrete cosine transform. *Int. J. Comput. Vision*, 43:167–188, July 2001.

[HM09] O.C. Hamsici and A.M. Martinez. Active appearance models with rotation invariant kernels. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1003 –1009, October 2009.

[HO00] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Netw.*, 13:411–430, May 2000.

[HP09] Abdenour Hadid and Matti Pietikaeinen. Manifold learning for gender classification from face sequences. In *Advances in Biometrics*, pages 82–91. Springer Verlag., 2009.

[HS88] C. Harris and M. Stephens. A Combined Corner and Edge Detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[JDN04] AK Jain, S. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. *Proceedings of International Conference on Biometric Authentication*, 3072:731–738, July 2004.

[KB01a] Timor Kadir and Michael Brady. Saliency, scale and image description. *Int. J. Comput. Vision*, 45:83–105, November 2001.

[KB01b] Timor Kadir and Michael Brady. Saliency, Scale and Image Description. *International Journal of Computer Vision*, 45(2):83–105, November 2001.

[KCT00] T. Kanade, J.F. Cohn, and Y Tian. Comprehensive database for facial expression analysis. *Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53, 2000.

[Kha03] S.A Khayam. The discrete cosine transform (dct): Theory and application. In *Michigan State University*, 2003.

[KWT88] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *INTERNATIONAL JOURNAL OF COMPUTER VISION*, 1(4):321–331, 1988.

[KZP08] V. Kellokumpu, G.Y. Zhao, and M. Pietikainen. Human activity recognition using a dynamic texture based method. pages xx–yy, 2008.

[LDC04] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. 34(1):621–628, February 2004.

[Lew95] J. P. Lewis. Fast normalized cross-correlation. In *Vision Interface*, pages 120–123. Canadian Image Processing and Pattern Recognition Society, 1995.

[LHK05] K.C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.

[LJ05] Stan Z Li and Anil K Jain. *Handbook of Face Recognition*. Springer, 2005.

[LK09] H.S. Lee and D.J. Kim. Tensor-based aam with continuous variation estimation: Application to variation-robust face recognition. *PAMI*, 31(6):1102–1116, June 2009.

[LLS09] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions from videos in the wild. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.

[LMS+] Ivan Laptev, Marcin Marszaek, Cordelia Schmid, Benjamin Rozenfeld, Inria Rennes, Irisa Inria Grenoble, and Lear Ljk. Learning realistic human actions from movies. In *In: CVPR. (2008*.

[LMT+07] Jun Luo, Y. Ma, E. Takikawa, S. Lao, M. Kawade, and Bao-Liang Lu. Person-specific sift features for face recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 2, pages II–593 –II–596, 2007.

[Low99] David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society.

[Low03] David G. Lowe. Distinctive image features from scale-invariant keypoints, 2003.

[LTC95a] A. Lanitis, C. J. Taylor, and T. F. Cootes. A unified approach to coding and interpreting face images. In *Proceedings of the Fifth International Conference on Computer Vision*, ICCV '95, pages 368–, Washington, DC, USA, 1995. IEEE Computer Society.

[LTC95b] A. Lanitis, C. J. Taylor, and T. F. Cootes. A unified approach to coding and interpreting face images. In *Proceedings of the Fifth International Conference on Computer Vision*, ICCV '95, pages 368–, Washington, DC, USA, 1995. IEEE Computer Society.

[LW99] Chengjun Liu and Harry Wechsler. Comparative assessment of independent component analysis (ica) for face recognition. In *International Conference on Audio and Video Based Biometric Person Authentication*, pages 22–24, 1999.

[LYS09] Jingen Liu, Yang Yang, and M. Shah. Learning semantic visual vocabularies using diffusion distance. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:461–468, 2009.

[May07] Christoph Mayer. Rigid 3d model. In *Workshop on Vision, Modeling, and Visualization (VMV)*, volume 1, pages 233–241, Saarbrücken, Germany, November 2007.

[MB03] Iain Matthews and Simon Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60:135–164, 2003.

[MGPW05] Raphael Marée, Pierre Geurts, Justus Piater, and Louis Wehenkel. Random subwindows for robust image classification. In *In CVPR*, pages 34–40. IEEE, 2005.

[MHS09] Pyry Matikainen, Martial Hebert, and Rahul Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. In *Workshop on Video-Oriented Object and Event Classification, ICCV 2009*, September 2009.

[MHS10] Pyry Matikainen, Martial Hebert, and Rahul Sukthankar. Representing pairwise spatial and temporal relations for action recognition. In *European Conference on Computer Vision (ECCV)*, September 2010.

[Mic] Microsoft. Kinect for xbox 360. http://www.xbox.com/en-US/kinect.

[MN08] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. *ECCV*, 2008. http://www.milbo.users.sonic.net/stasm.

[MPK09] Ross Messing, Chris Pal, and Henry Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV '09: Proceedings of the Twelfth IEEE International Conference on Computer Vision*, Washington, DC, USA, 2009. IEEE Computer Society.

[MR10] Christoph Mayer and Bernd Radig. Learning displacement experts from multiband images for face model fitting. In *2010 Third International Conferences on Advances in Computer-Human Interactions*. IEEE Computer Society, 2010.

[MSH+10] Yadong Mu, Ju Sun, Tony X. Han, Loong-Fah Cheong, and Shuicheng Yan. Randomized locality sensitive vocabularies for bag-of-features model. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, volume 6313 of *Lecture Notes in Computer Science*, pages 748–761. Springer, 2010.

[MSV+09] LM Maat, RC Sondak, MF Valstar, M Pantic, and P Gaia. Man machine interaction (mmi) database, 2009.

[MW09] David Matsumoto and Bob Willingham. Spontaneous facial expressions of emotion of congenitally and noncongenitally blind individuals. *Journal of Personality and Social Psychology*, 96(1):1–10, Jan 2009.

[MWER09] Christoph Mayer, Matthias Wimmer, Martin Eggers, and Bernd Radig. Facial expressions recognition with 3d deformable models. In *2009 Second International Conferences on Advances in Computer-Human Interactions*, pages 26–31. IEEE Computer Society, 2009.

[MWR09] Christoph Mayer, Matthias Wimmer, and Bernd Radig. Adjusted pixel features for facial component classification. *Image and Vision Computing Journal*, 2009.

[MWS+08] Christoph Mayer, Matthias Wimmer, Freek Stulp, Zahid Riaz, Anton Roth, Martin Eggers, and Bernd Radig. A real time system for model-based interpretation of the dynamics of facial expressions. In *Proc. of the International Conference on Automatic Face and Gesture Recognition (FGR08)*, Amsterdam, Netherlands, September 2008.

[Nef02] A.V. Nefian. Embedded bayesian networks for face recognition. *International Conference on Multimedia and Expo*, 2:133–136, 2002.

[NIH98] A.V. Nefian and M.H. III Hayes. Hidden markov models for face recognition. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal*, 5:2721–2724, 1998.

[NIH00] A.V. Nefian and M.H III. Hayes. Maximum likelihood training of the embedded hmm for face detection and recognition. *Proceedings of International Conference on Image Processing,*, 1:33–36, 2000.

[NS08] K. Netzell and J.E. Solem. Efficient image inner products applied to active appearance models. pages 1–4, 2008.

[OBSC00] Atsuyuki Okabe, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu. *Spatial tessellations: Concepts and applications of Voronoi diagrams*. Probability and Statistics. Wiley, NYC, 2nd edition, 2000. 671 pages.

[OPH96] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59, January 1996.

[orl] At & t the database of faces. http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.htm

[O'T09] Alice J O'Toole. Cognitive and computational approaches to face recognition. *The University of Texas at Dallas*, November 2009.

[PJ07] Unsang Park and Anil Jain. 3d model-based face recognition in video. In Seong-Whan Lee and Stan Li, editors, *Advances in Biometrics*, volume 4642 of *Lecture Notes in Computer Science*, pages 1085–1094. Springer Berlin/Heidelberg, 2007.

[PZVkC] In Kyu Park, Hui Zhang, Vladimir Vezhnevets, and Heui keun Choh. H.k.: Image-based photorealistic 3-d face modeling. *In: FGR*, 2004:49–56.

[RBMR08] K. Ramnath, S. Baker, I. Matthews, and D. Ramanan. Increasing the density of active appearance models. pages 1–8, 2008.

[RGBR09] Zahid Riaz, Suat Gedikli, Michael Beetz, and Bernd Radig. A unified features approach to human face image analysis and interpretation. In *Affective Computing and Intelligent Interaction, Amsterdam, Netherlands*. IEEE, 2009. Doctoral Consortium Paper.

[RGBR10] Zahid Riaz, Suat Gedikli, Michael Beetz, and Bernd Radig. 3d face modeling for multi-feature extraction for intelligent systems. *Computer Vision for Multimedia Applications: Methods and Solutions*, 1:73–89, October 2010.

[RHCL10] John Ruiz Hernandez, James Crowley, and Augustin Lux. "how old are you?" : Age estimation with tensors of binary gaussian receptive maps. In *Proceedings of the British Machine Vision Conference*, pages 6.1–6.11. BMVA Press, 2010.

[Ria04] Zahid Riaz. Development of a face recognition system for identification. Master's thesis, Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan, September 2004.

[Ria08] Zahid Riaz. Face recognition in coffee break scenarios. International Workshop of Cognition for Technical Systems, October 2008.

[Ria09] Zahid Riaz. Face recognition in coffee break scenarios. International Workshop of Cognition for Technical Systems, 2009.

[Ria10] Zahid Riaz. Face recognition in coffee break scenarios. International Workshop of Cognition for Technical Systems, September 2010.

[RLS09] Kishore K Reddy, Jingen Liu, and Mubarak Shah. Incremental action recognition using feature-tree. In *International Conference on Computer Vision*, 2009.

[RMBR09a] Zahid Riaz, Christoph Mayer, Michael Beetz, and Bernd Radig. 3d model for face recognition across facial expressions. In *Biometric ID Management and Multimodal Communication, Madrid, Spain*. Springer, 2009.

[RMBR09b] Zahid Riaz, Christoph Mayer, Michael Beetz, and Bernd Radig. Facial expressions recognition from image sequences. In *2nd International Conference on Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, Prague, Czech Republic*. Springer, 2009.

[RMBR09c] Zahid Riaz, Christoph Mayer, Michael Beetz, and Bernd Radig. Model based analysis of face images for facial feature extraction. In *Computer Analysis of Images and Patterns, Munster, Germany*. Springer, 2009.

[RMBR09d] Zahid Riaz, Christoph Mayer, Micheal Beetz, and Bernd Radig. Model based analysis of face images for facial feature extraction. *Computer Analysis of Images and Pattern*, pages 99–106, September 2009.

[RMW+09] Zahid Riaz, Christoph Mayer, Matthias Wimmer, Michael Beetz, and Bernd Radig. A model based approach for expression invariant face recognition. In *3rd International Conference on Biometrics, Alghero Italy*. Springer, 2009.

[RMWR08] Zahid Riaz, Christoph Mayer, Matthias Wimmer, and Bernd Radig. Model based face recognition across facial expressions. In *Journal of Information and Communication Technology*, December 2008.

[Rob] RobotCub. icub: An open source robot. http://www.robotcub.org/.

[Rom05] S. Romdhani. *Face Image Analysis using a Multiple Feature Fitting Strategy*. PhD thesis, Computer Science Department, University of Basel, Basel Switzerland, 2005.

[RS10] Mikkel Rodriguez and Mubarak Shah. Sprts database. 2010.

[RV03] Sami Romdhani and Thomas Vetter. Efficient, robust and accurate fitting of a 3d morphable model. In *IEEE International conference on Computer Vision 2003*, 2003.

[Ryd87] M. Rydfalk. *Candide, a parameterized face*. PhD thesis, Dept. of Electrical Engineering, Linkoeping University, Sweden, 1987.

[SAS07] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia*, pages 357–360, New York, NY, USA, 2007. ACM.

[SBOR06] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of IEEE*, 94(II):1948–1962, November 2006.

[SCH09] Xinghua Sun, Mingyu Chen, and A. Hauptmann. Action recognition via local descriptors and holistic features. *Computer Vision and Pattern Recognition Workshop*, 0:58–65, 2009.

[SGM09] Caifeng Shan, Shaogang Gong, and Peter McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. In *Computer Vision and Image Understanding*, pages 803–816. Elsevier Inc., 2009.

[SH08] M. Saquib Sarfraz and Olaf Hellwich. Head pose estimation in face recognition across pose scenarios. In *VISAPP (1)*, pages 235–242, 2008.

[Sha08] *Recognizing human actions using multiple features*, June 2008.

[SHN] Pascal Steingrube, Harald Hanselmann, and Hermann Ney. Dreuw et al.: Surf-face recognition 1 surf-face: Face recognition under viewpoint consistency constraints.

[SHR10] M. Saquib Sarfraz, Olaf Hellwich, and Zahid Riaz. Feature extraction and representation for face recognition. *Face Recognition, Milos Oravec (Ed.)*, 2010.

[SK87] L. Sirovich and M. Kirby. Low dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America*, 4(3):519–524, 1987.

[SLC04] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local svm approach. In *In Proc. ICPR*, pages 32–36, 2004.

[SMB00] Cordelia Schmid, Roger Mohr, and Christian Bauckhage. Evaluation of interest point detectors, 2000.

[SRE+05] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering Object Categories in Image Collections. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2005.

[Ste03] M. B. Stegmann. The AAM-API: An open source active appearance model implementation, nov 2003.

[Ste04] M. B. Stegmann. Generative interpretation of medical images, 2004.

[Sys] Intelligent Autonomous Systems. Assitive robotics for everyday life activities. http://ias.in.tum.de/research-areas/robots.

[TBB02] Sim Terence, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *FGR '02: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, page 53, Washington, DC, USA, 2002. IEEE Computer Society.

[TBB09] Moritz Tenorth, Jan Bandouch, and Michael Beetz. The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *IEEE Int. Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS). In conjunction with ICCV2009*, 2009.

[Tey08] Alexandra Teynor. *Visual object class recognition using local descriptions*. PhD thesis, Computer Science Department, Albert Ludwigs University, Freiburg Germany, 2008.

[TP91a] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3:71–86, January 1991.

[TP91b] Matthew Turk and Alex Pentland. Face recognition using eigenfaces. In *Proceedings of Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[VGM08] Kellokumpu V., Zhao G., and Pietikäinen M. Human activity recognition using a dynamic texture based method. In *Proc. The British Machine Vision Conference (BMVC 2008), Leeds, UK, 10 p*, 2008.

[VJ04] Paul Viola and Micheal J Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004.

[VLK+07] C. Vogler, Zhiguo Li, A. Kanaujia, S. Goldenstein, and D. Metaxas. The best of both worlds: Combining 3d deformable models with active shape models. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1 –7, 2007.

[VnU03] Michel Vidal-naquet and Shimon Ullman. Object recognition with informative features and linear classification. In *In ICCV*, pages 281–288, 2003.

[VS02] Julia Vogel and Bernt Schiele. On performance characterization and optimization for image retrieval. In *7 th European Conference on Computer Vision*, pages 49–63. Springer, 2002.

[Wel91] B. Welsh. *Model-Based Coding of Images*. PhD thesis, British Telecom Research Lab, 1991.

[WF05] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.

[WFKvdM97] Laurenz Wiskott, Jean-Marc Fellous, Norbert Krüger, and Christopher von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19:775–779, July 1997.

[WFS+08] Matthias Wimmer, Shinya Fujie, Freek Stulp, Tetsunori Kobayashi, and Bernd Radig. An ASM fitting method based on machine learning that provides a robust parameter initialization for AAM fitting. In *Proc. of the International Conference on Automatic Face and Gesture Recognition (FGR08)*, Amsterdam, Netherlands, September 2008.

[WH04] Z. Wen and T.S. Huang. *3D Face Processing: Modeling, Analysis and Synthesis*. Kluwer, 2004.

[WHS08] A. Wimmer, J. Hornegger, and G. Soza. Implicit active shape model employing boundary classifier. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1 –4, 2008.

[Wim07] Matthias Wimmer. *Model-based Image Interpretation with Application to Facial Expression Recognition*. PhD thesis, Technical University of Munich, October 2007.

[WMSR07] Matthias Wimmer, Christoph Mayer, Freek Stulp, and Bernd Radig. Estimating natural activity by fitting 3D models via learned objective functions. In *Workshop on Vision, Modeling, and Visualization (VMV)*, volume 1, pages 233–241, Saarbrücken, Germany, November 2007.

[WMY+08] John Wright, Student Member, Allen Y. Yang, Arvind Ganesh, Student Member, S. Shankar Sastry, Yi Ma, and Senior Member. Robust face recognition via sparse representation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI*, 2008.

[WRB10] Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. Research Report 7212, INRIA, February 2010.

[WRMR08] Matthias Wimmer, Zahid Riaz, Christoph Mayer, and Bernd Radig. Recognizing facial expressions using model-based image interpretation. *Advances in Human-Computer Interaction*, 1:587–600, October 2008.

[WSPR08] Matthias Wimmer, Freek Stulp, Sylvia Pietzsch, and Bernd Radig. Learning local objective functions for robust face model fitting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(8):1357–1370, 2008.

[WSTR06] Matthias Wimmer, Freek Stulp, Stephan Tschechne, and Bernd Radig. Learning robust objective functions for model fitting in image understanding applications. In Michael J. Chantler, Emanuel Trucco, and Robert B. Fisher, editors, *Proceedings of the 17th British Machine Vision Conference (BMVC)*, volume 3, pages 1159–1168, Edinburgh, UK, September 2006. BMVA.

[XBMK04] Jing Xiao, Simon Baker, Iain Matthews, and Takeo Kanade. Real-time combined 2d+3d active appearance models. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 535–542, 2004.

[Xim] Ximea. Currera-r series, smart camera with pc inside. http://www.ximea.com/en/products/intelligent-vision-systems/currera-r-series.

[ZC05] W.Y. Zhao and R. Chellappa. *Face Processing: Advanced Modeling and Methods*. Elsevier, 2005.

[ZCPR03] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey, December 2003.