Towards Automatic Face Recognition in Unconstrained Scenarios

von
Muhammad Saquib Sarfraz

Von der Fakultät IV- Elektrotechnik und Informatik
der Technische Universität Berlin
zur Erlangung des akadamischen Grades

Doktor der Ingenieurwissenschaften

-Dr.-Ing.-

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Opper
Gutachter:	Prof. Dr.-Ing. Olaf Hellwich
Gutachter:	Prof. Dr. Felix Wichmann

Tag der wissenschaftliche Aussprache: 30 September 2008

Berlin 2008
D 83

Zusammenfassung

Gesichtserkennung als aktiver Forschungsbereich der letzten zwei Jahrzehnte stellt immer noch viele Herausforderungen dar. Aktuelle Gesichtserkennungs-Systeme liefern nur befriedigende Ergebnisse unter kontrollierten Bedingungen. Die Erkennungsgenauigkeit lässt signifikant nach, wenn sie mit Änderungen von Blickwinkel, Beleuchtung und Fehlausrichtung konfrontiert werden. Das Hauptziel dieser Dissertation ist das Erforschen und Entwickeln neuer Methoden für ein vollautomatisches Gesichtserkennungs-System, welches in unkontrollierten Umgebungen arbeiten kann.

Im ersten Teil wird eine Merkmalsbeschreibung eingeführt, die robuster als ein pixelbasierter Ansatz ist. Sie ist invariant bezüglich Fehlausrichtung bei nicht perfekter Lokalisierung der Gesichter. Für die Mehrbild-Gesichtserkennung wird ein vollständiger Leistungsvergleich verschiedener Klassifikatoren in unterschiedlichen Merkmalsräumen präsentiert. Viele neue Ansätze befürworten die Berechnung von künstlichen Bildern aus verschiedenen Ansichten für ein gegebenes Gesichtsbild, um eine betrachtungsinvariante Erkennung zu realisieren. In ausführlichen Experimenten wird die Schwäche existierender Gesichtserkennungs-Systeme für kleine Mustergrößen demonstriert. Um die Erkennung zu verbessern, wird ein Schema zur Kombination von Klassifikatoren für verschiedene Merkmalsräume vorgeschlagen.

Im zweiten Teil der Arbeit stellen wir ein neues System zur Schätzung der Kopfhaltung vor. Die Blickwinkelinformation ist nützlich und für ein vollautomatisches Gesichtserkennungs-System muss die Ansicht in den Eingangsbildern bekannt sein. Das vorgeschlagene Blickwinkelschätzungs-System funktioniert bei hohen Beleuchtungs- und Ausdrucksänderungen. In diesem Zusammenhang haben wir eine neue Merkmalbeschreibung mit dem Namen LESH eingeführt, welche die zugrunde liegende Form beinhaltet und unempfindlich bezüglich der Hautfarbe und verschiedener Beleuchtungen ist. Basierend auf der vorgeschlagenen LESH Merkmalsbeschreibung, wird ein generischer Ähnlichkeitsraum generiert, welcher nicht nur eine effektive Dimensionalitäts-Reduzierung sondern auch viele repräsentative Vektoren für einen bestimmten Testmerkmalsvektor bietet. Dieser wird verwendet, um Wahrscheinlichkeiten für verschiedene Blickwinkel zu generieren, ohne explizit die zugrunde liegende Dichte zu schätzen, was sehr nützlich für die nachfolgende Gesichtserkennung unter verschiedenen Blickwinkeln ist.

Im dritten Teil dieser Dissertation integrieren wir das System zur Schätzung der Kopfhaltung mit einem neuen vollautomatischen Gesichtserkennungs-System. Wir stellen eine betrachtungsinvariante Gesichtserkennungsmethode vor, welche nur ein Einzelbild der Person aus einer Galerie zur Erkennung benötigt. Der vorgeschlagene Ansatz konzentriert sich auf das Modellieren von Verbundansichten der Galerie und der Testbilder über verschiedene Ansichten in einem Bayesschen Ansatz. Diese Methode liefert einen vollständigen Posterior über alle möglichen Galeriezuordnungen, welcher auch einfach für Gesichts-Authentifikation genutzt werden kann. Unsere Methode benötigt keine strikte Ausrichtung zwischen der Galerie und dem Testbild, was es verglichen mit den momentanen Methoden besonders attraktiv macht. Die vorgeschlagenen Algorithmen wurden mit mehreren Referenz-Datenbanken ausgewertet, die tausende herausfordernder Bilder mit verschiedenen Variationen bezüglich Ausdruck, Ansicht und Beleuchtung enthalten. Die Ergebnisse zeigen, dass unsere Methoden eine deutliche Verbesserung gegenüber bisherigen Ansätzen darstellen.

Abstract

Face recognition, as an active area of research over the past two decades, still poses many challenges. Current face recognition systems yield satisfactory performance only under controlled scenarios and recognition accuracy degrades significantly when confronted with unconstrained situations due to variations such as pose, illumination and misalignments etc. The principal objective of this dissertation is to investigate and introduce new methods towards building a fully automatic face recognition system that can work in unconstrained environments.

In the first part we introduce to use a more robust feature description than pixel based appearances that is invariant with respect to misalignments due to non-perfect localization of faces. A thorough performance analysis of many different classifiers on different feature spaces in the context of multi-view face recognition is presented. Many recent approaches advocate the use of generating artificial images at different views for a given face image in order to realize pose invariant recognition. It is demonstrated in an extensive experimental setting the weakness and applicability of existing face recognition systems with respect to small sample size problem in these situations. Furthermore a classifier combining scheme over different feature spaces and different classifier is proposed to improve recognition.

In the second part of the thesis we present a novel head pose estimation system. The pose information is valuable and for a fully automatic face recognition system the pose of the incoming image has to be known. The proposed front-end pose estimation system functions in the presence of large illumination and expression changes. In this context we have introduced a new feature description termed as LESH, which encodes the underlying shape and is insensitive to skin color and illumination variations. Based on proposed LESH feature description, we introduced to generate a generic similarity feature space, that not only provides an effective way of dimensionality reduction but also provides us with many representative vectors for a given test feature vector. This is used in generating probability scores for each pose without explicitly estimating the underlying densities, which is very useful in later face recognition across pose scenarios.

Finally, in the third part of this dissertation we integrate the head pose estimation system with a novel fully automatic face recognition system. We introduce a pose invariant face recognition method that requires only single image of the person to be recognized, in the gallery. The proposed approach is centered on modeling joint appearance of gallery and probe images across pose in a Bayesian framework. The method provides us with a full posterior over possible gallery matches which can also be easily used for face authentication. Our method does not require any strict alignment between gallery and probe images and that makes it particularly attractive as compared to the existing state of the art methods. The proposed algorithms have been evaluated on a number of benchmark databases, which contain thousands of challenging images with different variations in expression, pose, and illumination. Results indicate that our methods make appreciable improvement over the previous state-of-the-art approaches.

Acknowledgments

All praises are for Allah, most gracious most merciful, the creator of all that exists. Certainly it is He who has enabled me today to achieve one of the most cherished dreams of my life, a PhD degree.

I would like to express my sincere gratitude to my supervisor Prof. Olaf Hellwich who has always been a fatherly figure for me in these past four years. I am indebted for his valuable guidance on both scientific and personal matters. I am also grateful to Prof. Lothar Gründig for always writing positive recommendations for my yearly funding extensions. I would also take this opportunity to thank my committee members Prof. Felix Wichmann and Prof. Opper for carefully reviewing my thesis.

I had a very pleasant stay at computer vision and remote sensing laboratory. I enjoyed my collaborations and discussions with my colleagues, Marc Jaeger, Adam Stanski, Hongwei Zheng, Oliver Gloger, Ulas Yilmaz, Ronny Hänsch, Mathias Heinrich, Anke Bellmann and Volker Rodehorst. I am especially beholden to my close collaborators Marc and Adam for all the fruitful scientific discussions and help they have been in all these years. This thesis would not have been possible without these brilliant guys. Thanks to Ms. Marion Dennert, our secretary, for always being helpful regarding administrative chores. Special thanks to Mathias, my office mate, for putting up with me and my weird jokes.

I would also like to thank Higher Education Commsion (HEC), Pakistan and Deutscher Akadamischer Austausch Dienst (DAAD), Germany for funding my PhD here at Technical University of Berlin.

I am mostly indebted to my wife Ayesha for her patience, her encouragement, and her lifelong love. She has made many sacrifices in order to enable me to complete this dissertation. This is her accomplishment as well. Last but not the least, thanks to my daughter Aimel for always putting up a smile on my face with her innocent ways in the most stressed times. Watching her grow makes me realize that there are few things in this world we do not want to automate.

This thesis is dedicated to my late mother Shehnaz. It was her dream that I pursue a scientific career. I wish she is with us today to watch me achieving an important milestone of my life.

Thank you All!

Table of Contents

List of Figures

List of Tables

List of Publications

Some parts of the work presented in this dissertation have been published in the following articles.

1. Book Chapter: M.S.Sarfraz and O.Hellwich, “On Head Pose Estimation in Face Recognition”, J. Braz, A. Ranchordas, H. Araújo and J. Jorge Editors, Advances in Computer Graphics and Computer Vision. Springer, 2009. (To appear).

2. M.S.Sarfraz, O.Hellwich, “Statistical Appearance Models for Automatic Pose Invariant Face Recognition”, Image and Vision Computing Journal, Elsevier Science, 2009 (invited).

3. M.S.Sarfraz, O.Hellwich,” An efficient front-end facial Pose estimation system for Face Recognition”, Int. Journal of Pattern Recognition and Image Analysis, distributed by Springer, Vol.18 (3), pp. 434–441. 2008.

4. M.S.Sarfraz and O. Hellwich “Statistical Appearance Models for Automatic Pose Invariant Face Recognition.” 8th IEEE Int. conference on Face and Gesture recognition FG, Holland, 2008.

5. M.S.Sarfraz, O.Hellwich ,"Head Pose Estimation in Face Recognition across Pose Scenarios" in Int. conference on computer vision theory and applications VISAPP, Vol. 1, pp. 235 -242, Portugal, 2008. (**Best Paper Award**).

6. M.S.Sarfraz , O.Hellwich, "Learning Probablistic Models for Recognizing faces under Pose Variations" in VISAPP-IMTA Workshop Image mining, theory and applications, pp. 122-132, Portugal, 2008.

7. M. S. Sarfraz, O. Hellwich.:"Robust Facial Pose Estimation for Face Recognition", In Proceedings of 7th IAPR Open German Russian Workshop on Pattern Recognition and Image Analysis, Germany, 2007.

8. M.S.Khalid, M.U.Ilyas, M.S.Sarfraz, M.A.Ijaz,” Bhattacharyya Coefficient in Correlation of Gray Scale Objects”, In Journal of Multimedia, Academy publishers, Vol. 1, No. 1, pp. 56-61, 2006.

9. M.S.Sarfraz, M.Jäger, O.Hellwich, “Performance Analysis of Classifiers on Face Recognition”, 5th IEEE Advances in Cybernetics Systems AICS conference, pp. 255 -264, United Kingdom, 2006.

10. M.S.Khalid, M.B.Malik, M.U.Ilyas, M.S.Sarfraz, K.Mahmood, “Performance of a Similarity Measure in Grayscale Image Matching” IEEE International Conference of Emerging Technologies, ICET, pp 121-125, Pakistan, 2005.

PART – 1

Face Recognition: Introduction and Current Challenges

Chapter 1. Introduction

Identifying persons by looking at their faces is a trivial task for humans. A baby is able to recognize his mother after few hours of his birth. Since the advent of digital computers, tireless attempts have been made to give them our most complicated sense, vision. The significant progress in this direction reflects the ability of current machine vision systems to work reliably in variable environments. Most systems are able to perform a number of vision tasks reliably e.g. recognizing and categorizing natural objects in scenes, advanced vision guided systems are able to perform on practical situations such as autonomous vehicle guidance, optical character recognition systems and so on. But when it comes to identifying persons based on recognizing their faces, what seems to be an overly simple task for humans becomes increasingly difficult and challenging for computers. With the increasing demands put on privacy, security and surveillance and an increasing range of applications needing a reliable verification system [OBRG04] [L02] [A01], automatic person identification has become very important topic for the research community. Traditional knowledge-based (password or Personal Identification Number 'PIN') and token-based (passport, driver license, and ID card) identifications are prone to fraud because PIN may be forgotten or guessed by an impostor and the tokens may be lost or stolen. A reliable identity authentication system will therefore need a biometric component. Among biometric metrics e.g. finger print, iris, face and palm, face is the most important metric to be utilized because of its non-intrusive nature as little or no cooperation is needed from the users in order to capture his/her face.

Over the past two decades several attempts have been made to address this problem and a voluminous literature has been produced. Current face recognition systems are able to perform very well in controlled environments e.g. frontal face recognition, where face images are acquired under frontal pose with strict constraints as defined in related

face recognition standards [ANSI04]. However, in unconstrained situations where a face may be captured in outdoor environments, under arbitrary illumination and large pose variations these systems fail to work. Nonetheless, due to the very demanding nature of the problem, there is a need to overcome these constraints and build face recognition systems that can recognize faces reliably even when captured under unconstrained scenarios. DARPA in United States has initiated face recognition technology evaluation test protocol 'FERET' in order to record and streamline the work in this direction. Face recognition grand challenge FRGC and face recognition vendor test FRVT are two most important events held every few years in order to evaluate the progress for unconstrained face recognition. While many systems and approaches are being documented, recognizing faces under large pose variations and illumination conditions is still the biggest problem as documented by recent FRGC and FRVT reports [FRVT02] [PFSBW06].

This dissertation attempts to address the problem of unconstrained face recognition and contributes towards building a fully automatic face recognition system that is able to recognize faces under large pose and illumination variations. Let us first briefly survey some of the basic concepts in automatic face recognition.

1.1 Automatic Face Recognition

A general face recognition system is comprised of different components, Figure 1.1. The main parts are typically face detection and face recognition that can be further decomposed in normalization, feature extraction and classification steps. Face detection is the very first important element in any automated face recognition system. Given an arbitrary image or image sequence, the goal of a face detection system is to determine the presence of a face in this image as well as its location. If a face is present, the system returns the location and its extent. Most of the fully automatic face recognition systems presented in the literature involve a separate face detection and/or localization step. Numerous approaches have been proposed to tackle the problems of face detection, see [YKA02] for a recent survey. We can differentiate face detection from face localization.

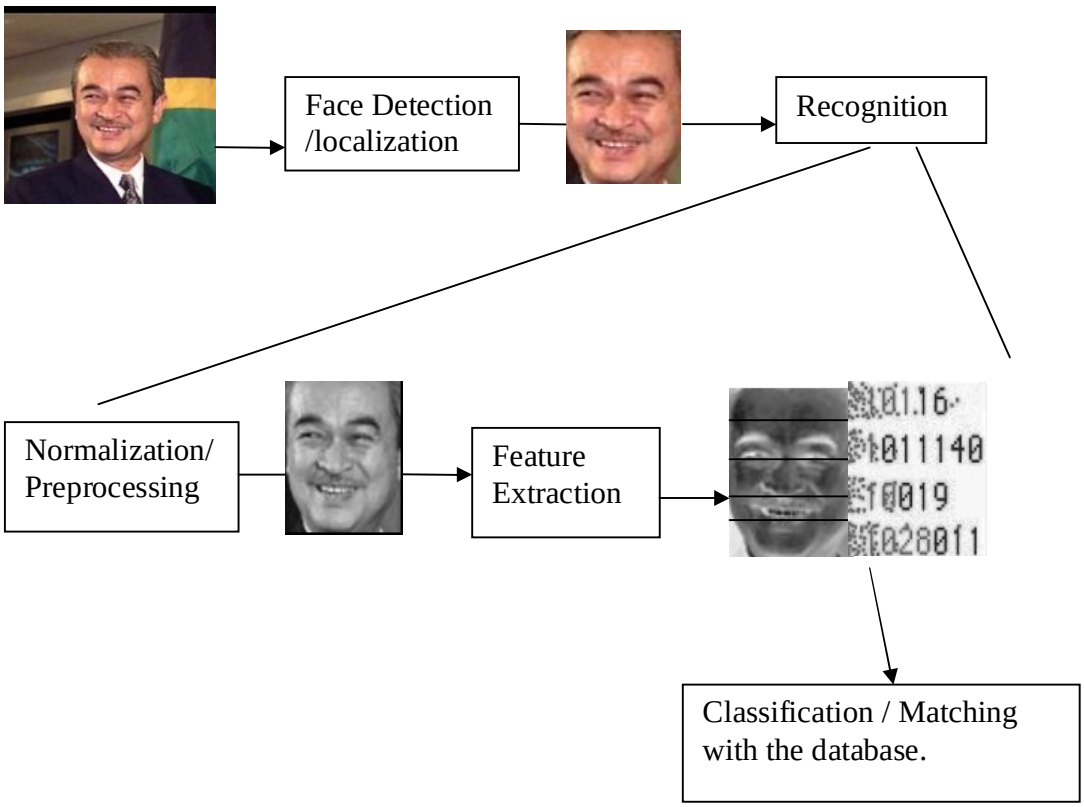


Figure 1-1: Main Components of a General Automatic Face Recognition System

Face detection aims to determine whether or not there are any faces in the image and, if present, returns the face location while the goal of face localization is to estimate the position of a single face. Localization usually needs to provide a more precise position of the face than detection and can be more difficult in this case.

After a detection and localization stage the face is subject to normalization. Normalization or preprocessing usually involves geometric normalization for the purpose of alignment and preprocessing for the illumination compensation. The alignment means detecting some landmark points on the face and warping the image onto a plane where these points are at some fixed locations. Normalized face images are then used for extracting features. The features are the meaningful information from the

image. These features are then modeled and finally matched by a classification or matching stage.

The purpose of face recognition as a whole is to identify or verify a face against a stored database of images. It mainly involves the following three tasks, adapted and quoting from [FRVT02].

- Verification. The recognition system determines if the query face image and the claimed identity match.
- Identification. The recognition system determines the identity of the query face image by matching it with a database of images with known identities, assuming that the identity is inside the database.
- Watch list. The recognition system first determines if the identity of the query face image is on the stored watch list and, if yes, then identifies the individual.

The main focus of this dissertation is the task of face identification. In many of the previous published literatures, the term 'face recognition' is used with the meaning of 'face identification'. In this dissertation the task of face identification is referred to as face recognition as well.

1.2 The Problem

Recognition of human faces, which are three-dimensional '3D' deformable objects, from their two-dimensional '2D' images poses many challenges. Today, several systems that achieve high recognition rates have been developed, however, such systems work in controlled environments. For most of them, face images must be frontal or profile, background must be uniform and lighting must be constant. Furthermore, lots of published systems are evaluated using manually located faces and the ones which have been evaluated using a fully automatic system showed a big degradation in performances [PFSBW06]. In most real life applications, the environment is not known *a-priori* and the system should be fully automatic. An unconstrained face recognition system has to deal with the following problems:

- Head Pose changes

- Lighting Variation
- Non-Perfect Detection
- Alignment/ imperfect localization of faces
- Occlusion, aging etc.

In this thesis, we will focus on face recognition approaches towards robust face recognition in unconstrained environments. Here, unconstrained means that the illumination of the face, the head position and the background are not known *a priori* and can change from the reference images to the probe images. In this context, the face recognition task presents new difficulties such as large variability in the images for the same identity, the lack of reference images and face alignment problems. To our knowledge, no existing face recognition system can combine accuracy and robustness in unconstrained environment.

The main problem of unconstrained face recognition is the large variability between face images. In fact the variations of appearance present in the images of the same person due to different lighting, pose etc are much larger than the appearance variations present between the images of two different persons in same lighting and/or pose. Pose change appear when the face changes its position and orientation in 3D space relative to the camera. However, a face can also undergo non-rigid motion when its 3D shape changes due to other factors such as speech or facial expression.

In order to cater these in-class variations, it is commonly believed that if a large training set is available comprising of all the different images representing these variations for each person, one can increase the robustness of the recognition system by modeling explicitly these in-class variations for each person. However, collecting many images of the same person covering, for instance, wide range of head pose changes, different illumination and expressions is very difficult , costly and not practical under realistic conditions. Despite this fact, most of the existing security and surveillance applications generally can store only one or few images of a particular person of interest. Such applications are e.g. police criminal record databases (mug-shots), border control, passport control etc. The problem of having very few training images available for each person is referred to as small sample size problem in machine learning literature. For example from a statistical stand point, number of training images should be 10 times as

that of the dimensionality of the problem. [DHS01]. That implies for a 100x100 face image, represented as a point in some feature space, by e.g. concatenating all the pixels into a vector, will yield a 10000 dimensional feature space. And theoretically, therefore we need 100000 training images for that person. On the other hand, face recognition is different from other machine learning tasks in that, here each person defines a unique class, and by the very nature of the problem where one has to distinguish among thousands of persons, we generally have a plethora of classes. Storing several images for each person also puts strong constraints on memory and storage among other factors. From a more practical stand point, therefore, it is generally needed to have a system that can recognize a person by having seen his/her only one image. Recognizing a person reliably having seen only one image and from previously unseen view point is a very challenging and unsolved problem.

1.3 Goals and Contributions of this Work

In this section we define the main goals and summarize the major contributions of this thesis.

1.3.1 Main goals

The thesis aims to develop a fully automatic face recognition system that works well under large 3D pose variations and requires only one 2D training image per person in the database. While true 3D based approaches in theory allow face matching at various poses, current 3D sensing hardware has too many limitations [BCF06], including cost and range. Moreover unlike 2D recognition, 3D technology cannot be retrofitted to many of the existing systems and applications such as surveillance systems, video feeds, criminal records; access control etc. 2D face recognition methods are therefore needed to be further investigated in order to generalize well under unconstrained scenarios.

Previous approaches addressing pose variations include the synthesis of new images at previously unseen views [BGPV05] [CSCG07], or direct synthesis of face model parameters [CSB06]. No matter how one approach to address this problem, the pose of the incoming test probe image has to be known a priori in order to realize a fully

automatic face recognition system. An efficient front-end head pose estimation system is, therefore, needed for any automatic pose invariant face recognition system to work. A fully functional head-pose estimation system that may be easily integrable into any of the existing face recognition systems is also one of the goals of this thesis.

Furthermore the 'one training image per person' problem needs to be carefully addressed. The one sample problem is defined as follows:

> "Given a stored database of faces with only one image per person, the goal is to identify a person from the database later in time in any different and unpredictable poses, lighting, etc. from just one image" [TCZZ06].

Given its challenge and significance for real-world applications, this problem is rapidly emerging as an active research area of face recognition. Effective algorithms that deal with this problem are also the goal of this dissertation.

Apart from pose variations, imperfect face localization [RCBM06] is also an important issue in a real life recognition system. Imperfect localizations or misalignments result in translations as well as scale changes, which adversely affect recognition performance. Finding invariant feature descriptions that can handle imperfect localizations and do not require strict alignment is also addressed.

1.3.2 Major contributions

The major contributions of this dissertation are as follows:

- A thorough performance analysis of many different classifiers on different feature spaces in the context of multi-view face recognition is presented. Many recent approaches advocate the use of generating artificial images at different views for a given face image or transform the incoming image to its frontal counter part in order to realize pose invariant recognition. It is demonstrated in an extensive experimental setting the weakness and applicability of existing face recognition systems with respect to small sample size problem in these situations. Furthermore a classifier combining scheme over different feature spaces and different classifier is proposed to improve recognition [SJH06].

- A feature description termed as face-GLOH signatures based on [MS05] has been introduced, for the task of face recognition for the first time, that do not require the face images to be properly aligned.
- A front-end head pose estimation system has been developed, that works in the presence of large illumination changes and is invariant to person specific appearance variations [SH07] [SH09a].
- A novel feature description for the purpose of pose estimation termed as LESH (Local Energy based Shape Histogram) is proposed. The new feature description primarily models the underlying shape and can be used in other similar computer vision tasks such as shape based image retrieval, object tracking, object recognition etc. [SH08c].
- A new classification procedure based on generating a generic multidimensional similarity feature space is proposed. The proposed approach is also useful for other similar machine vision tasks [SH08a].
- A novel fully automatic face recognition system based on probabilistic learning that works reliably with just one training image is developed and it is shown robust to large pose variations [SH08b].

The proposed algorithms have been evaluated on a number of benchmark databases, which contain thousands of challenging images with different variations in expression, pose, and illumination. Results indicate that our methods make appreciable improvement over the previous state-of-the-art approaches.

1.4 Organization of the Thesis

The thesis has been divided in three parts.

Part 1:

Chapter 1: General introduction

Chapter 2: Presents a brief survey on state of the art face recognition approaches and elaborate more on some of the underlying key elements of face recognition systems. The face databases and corresponding protocols used in our experiments are also introduced.

Chapter 3: Presents some of the most important commonly used feature extraction techniques and introduces a new feature description to be used in face recognition that is robust against misalignments and imperfect localization.

Chapter 4: Presents a thorough performance analysis of the common classifiers used in multi-view face recognition and explores the benefit of classifier combining. Furthermore, small sample size problem and its effects, with respect to imperfect localization of faces, large pose differences and large number of subjects are studied in an extensive experimental setup.

Part 2:

Chapter 5: Introduces the problem of head pose estimation and presents a novel feature description based on local energy model that encodes the underlying shape well.

Chapter 6: Introduces the new classification procedure to be used for estimating poses. Experimental results and comparison with previous state of the art methods is presented.

Part 3:

Chapter 7: Proposes a novel pose invariant face recognition approach that requires only single training image per person and works on misaligned images.

Chapter 8: Presents the experimental results and comparisons with previously published methods that demonstrate the effectiveness of the approach.

Chapter 9: Conclusions and future research directions.

Chapter 2. State of the Art in Face Recognition

The goal of this chapter is to provide some background and specifics of automatic face recognition. First the particulars of the preprocessing prerequisites of almost all of the current face recognition systems are described; the insight into the alignment and localization stage and its impact on the overall performance of face recognition system is highlighted, then a brief review of the most popular methods used in face recognition is provided and finally the description of the face databases and associated evaluation protocols used along the thesis are presented.

2.1 Face Recognition: Prerequisites

After a face has been detected in an image by a face detector, see a recent survey on face detection in [YKA02], the rest of the task of recognition can be categorized largely into the following three subtasks.

- Normalization
- Feature Extraction
- Classification

Normalization, also called preprocessing stage, is a prerequisite of any face recognition system and it generally involves geometric normalization and illumination normalization. Illumination normalization is used in order to compensate for large lighting variations and removing shadows etc, by using any of the lighting normalization methods, see [HCM05] for a comparison of different methods. Here we focus on the task of geometric normalization, also known as localization and alignment in the face processing jargon.

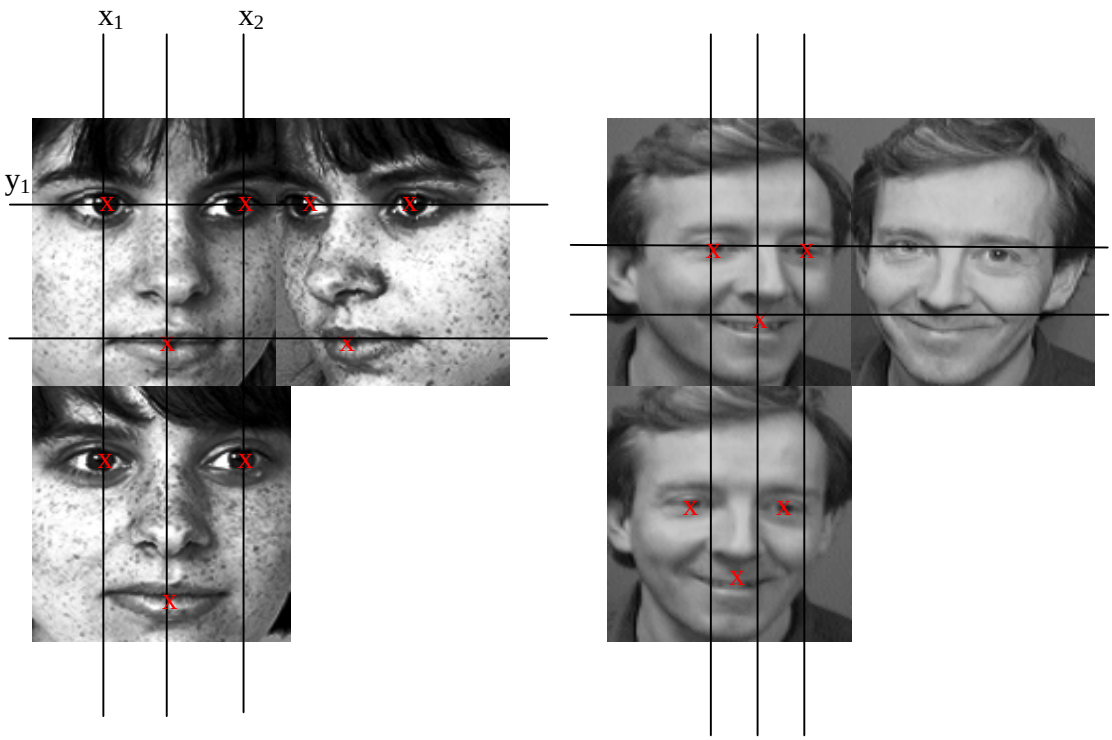


Figure 2-1: Difference between aligned and misaligned faces. (a) Aligned images with respect to three detected landmarks i.e. center of eyes and mouth position. (b) Localized faces without alignment

Note that the term localization and alignment is sometimes used interchangeably in the literature [M02], while here we differentiate between the two with respect to the strict role they play in later recognition stages.

The goal of face localization is to estimate the position of a single face. Localization usually needs to provide a more precise position of the face than detection.

Alignment involves the detection of some fiducial points on the face, such as facial landmarks e.g. center of eyes, nose tip etc., and fixing these points to predefined locations for all images. For instance, the eye centers, the medial line of the nose or mouth center etc., are expected to be at the same pixel coordinates in all images. To achieve that, special warping procedures are used [MT08]. This warping procedure is necessary to guarantee that every image pixel represents the same feature and thus to

achieve better recognition, this is known as the correspondence problem [BP96], [MK01].

2.1.1 Alignment of faces

Alignment itself has emerged as an important sub-problem in the face recognition literature [WTJ06], and a number of systems exist for the detailed alignment of specific facial parts [HFT03] [HLLYS04] [M05] [ZZTS05]. All face recognition algorithms require some degree of alignment so as to normalize for unwanted shape variations. Most of the algorithms use/advocate aligning images with respect to just 2-3 facial landmarks such as center of eyes, nose tip etc., faces aligned with 2-3 facial landmarks points are also called sparsely registered faces in the literature [LC08]. On the other hand some systems identify more than 80 landmarks (densely registered) per face. In a recent work Gross et al. [GMB04] demonstrated that improved face recognition performance can be attained using dense registration (39–54 fiducial points depending on the pose) rather than sparse registration (3 fiducial points located on the eyes and nose tip) especially for the task of pose-invariant face recognition. Similarly, Blanz and Vetter [BV03] demonstrated good performance using extremely dense offline registration (75,972 vertex points on laser scan 3D images) and medium density registration (at least 7–8 fiducial points depending on pose) with the online 2D images.

A problem with both these approaches, however, is that automatic dense registration of the face across viewpoints remains a very difficult task making most of these algorithms still very reliant on manual registration. Sparse registration (i.e., 2–3 fiducial points such as the eyes and nose) of the face is generally considered an easier problem than dense registration. Techniques for sparse registration are more mature than their denser counterpart, and can now perform very well on frontal faces (see [EZ06] for a review). However, automatically detecting several or even 2-3 facial landmarks on different views of the face is not a trivial task. None of the methods up till now can guarantee an acceptable performance in detecting even 2-3 points reliably across different views. This is largely due to the fact that when a face rotates from frontal to left or right profile view, the appearance of individual facial parts also change considerably

and some parts simply disappear, e.g. when a face moves from frontal to right profile the left eye will not be visible any more and hence, any methods that try to detect center of both eyes will eventually fail. Current algorithms addressing recognizing faces across pose variations, therefore, rely mostly on manually located landmarks for the purpose of alignment.

Keeping in view the preceding discussion, one may instantly note that this over reliance of almost all of the current face recognition systems on a strict alignment, either via sparsely registered faces or with densely registered faces, becomes counterproductive for a fully automatic face recognition system, especially in unconstrained scenarios where a large pose variation may be expected. This is because automatically detecting several facial landmarks with pixel accuracy can never be guaranteed, leading to misalignments which in turn has an adverse effect on recognition performance.

One way of overcoming this limitation for fully automatic face recognition systems would be to use large training sets for each person covering all possible variations of the faces due to misalignments and/or other variations, as also noted in [M02]. However, as indicated earlier, this is practically not achievable. The other possible way would be to investigate such facial representations which are inherently robust or invariant against such misalignments and therefore do not require a strict alignment procedure. We will focus on this goal in the next chapter.

2.2 Face Recognition: Literature Review

The development of machines capable of automatic face perception has been multi-disciplinary. It has benefited from areas as varied as computer science, cognitive science, mathematics, physics, psychology and neurobiology. For over three decades a voluminous literature on face recognition has been produced. Psychology has shown that the factors contributing to successful face recognition are quite complex. Nevertheless changes in facial expression and angle do affect recognition of unfamiliar faces whereas recognition of familiar faces is not affected by such changes. Moderate changes in viewing angle and facial expressions do not affect recognition accuracy. It is clear, therefore, that different processes are engaged in the recognition of familiar and

unfamiliar faces. Recognition of familiar faces appears to be relatively robust against changes in viewing angle, whereas unfamiliar faces are affected by rotation of the head [PA03].

Neurobiologists have long been studying the mechanism by which the brain recognizes faces, many of whom believe that the brain perceives faces as "special" and very different from other visual objects. For example, classic studies found that turning the image of a face upside down compromises recognition much more than it does for similarly inverting other objects. In a ground breaking work [MWB97] demonstrated that human mind process faces in a separate special area of the brain called *fusiform* that is separate from the general purpose visual processing system and as such recognizing faces is different from recognizing general objects. This has been an overwhelming view over the past decade until recently new studies [J06] provide a compelling array of evidence supporting the idea that the processing of faces and objects do not rely on qualitatively different mechanisms. Computer scientists are using these guidelines to enable machines to see and identify faces among other objects. The pattern recognition and computer vision techniques developed have largely been investigated to solve the problem of machine-based or computer based face recognition.

Computer-based face recognition can be mainly categorized into appearance-based methods and model-based methods. Appearance-based methods tend to describe the whole appearance of the face by vectorizing the face image pixels into a $\mathbb{R}^n$ feature space. Subspace analysis is then carried out into this feature space. Since the common appearance-based approaches builds on the image pixels, they are very sensitive to appearance variations caused due to e.g. lighting and pose. Model based approaches builds a person specific model of the face based on modeling selected features on the face. Below we review some of the most important methods employed in these two domains.

2.2.1 Appearance-based methods

Many approaches to object recognition and to computer graphics are based directly on images without the use of intermediate 3D models. Most of these techniques depend on a

representation of images that induces a vector space structure and, in principle, requires dense correspondence. Appearance-based approaches represent an object in terms of several object views (raw intensity images). An image is considered as a high-dimensional vector, i.e. a point in a high-dimensional vector space. Many view-based approaches use statistical techniques to analyze the distribution of the object image vectors in the vector space, and derive an efficient and effective representation (feature space) according to different applications. Given a test image, the similarity between the stored prototypes and the test view is then established in the feature space.

2.2.1.1 Subspace analysis

Image data can be represented as vectors, i.e., as points in a high dimensional vector space. For example, a $m \times n$ 2D image can be mapped to a vector $\mathbf{x} \in \mathbf{R}^{\mathbf{mn}}$ by lexicographic ordering of the pixel elements (such as by concatenating each row or column of the image). Despite this high-dimensional embedding, the natural constraints of the physical world (and the imaging process) dictate that the data will, in fact, lie in a lower-dimensional manifold. The primary goal of the subspace analysis is to identify, represent, and parameterize this manifold in accordance with some optimality criteria.

Principal component analysis 'PCA' is the most common way of finding the subspace that describes the most variance of the data. One of the benchmark appearance based method, Eigenface [TP91] uses PCA to construct this subspace also termed as face space. The aim of PCA is to identify a subspace spanned by the training images which could decorrelate the variance of pixel values. The PCA finds the orthogonal directions that account for the highest amount of variance. The data is then projected into the subspace spanned by these directions. In practice, the principal component axes are the eigenvectors of the covariance matrix of the data. The corresponding Eigen-values indicate the proportion of variance of the data projections along each direction. Thus for 'M' training images $\{x_1, x_2, \ldots, x_M\}$, where each x is a d-dimensional vector, one can find the principal components by Eigen analysis of the covariance matrix $\sum$.

$$\sum = \frac{1}{M-1} \sum_{i=1}^{M} (x_i - \mu)(x_i - \mu)^T \tag{2.1}$$

Figure 2-2: Eigenvectors corresponding to the 7 largest Eigen-values for a subject shown as images (derived from ORL face database [ORL]).

The largest Eigen-vectors are then found by using corresponding Eigen analysis,

$$\sum E = \Lambda E \tag{2.2}$$

where E and Λ are the resulting Eigen vectors, also known as Eigen faces, and Eigen values respectively. The representation of a face image in the PCA subspace is then obtained by projecting it to the coordinate system defined by the Eigen faces [TP91]. The Eigen-faces corresponding to the 7 largest Eigen values, derived from ORL face database [ORL], are shown in Fig. 2-2. Several extensions of PCA are developed, such as modular Eigen spaces [PMS94] and probabilistic subspaces [BM02].

PCA can achieve the optimal representation in the sense of maximizing the overall data variance. However, the difference between faces from the same person due to illumination and pose (within-class scatter) seems to be larger than that due to facial identity (between-class scatter). Based on this observation, linear discriminant analysis (LDA) is applied that uses the class information to select a subspace in which different classes are optimally represented. LDA uses Fisher face methods [BHK97]. The Fisher face algorithm is derived from the Fisher Linear Discriminant (FLD), which uses class specific information. By defining different classes with different statistics, the images in the learning set are divided into the corresponding classes. Then, techniques similar to those used in Eigen face algorithm are applied. The Fisher face algorithm results in a higher accuracy rate in recognizing faces when compared with Eigen face algorithm. LDA defines a projection that makes the within-class scatter small and the between class scatter large. This projection has shown to be able to improve classification performance over PCA. However, it requires a large training sample set for good generalization, which is usually not available for face recognition applications. To address such Small Sample Size (SSS) problems, Zhao et al. [ZKCSW98] perform PCA to reduce feature dimension before LDA projection.

By using higher order statistical analysis, Independent Component Analysis (ICA) was first adopted by [BMS02] for face recognition. The work showed that ICA outperformed PCA. However, other researchers [DBBB03] observed that when the right distance metric is used, PCA significantly outperforms ICA on the FERET database.

PCA, LDA and ICA are common linear subspace projections, however, since face data defines a non-linear manifold into the feature space, non-linear subspace analysis techniques have also been developed. Recently, kernel methods have been successfully applied to solve pattern recognition problems because of their capacity to handle nonlinear data. By mapping sample data to a higher dimensional feature space, effectively a nonlinear problem defined in the original image space is turned into a linear problem in the feature space [SS02]. PCA or LDA can subsequently be performed in this feature space and are thus called Kernel Principal Component Analysis (KPCA) and Generalized Discriminant Analysis (GDA) [BA00]. Experiments show that KPCA and GDA are able to extract nonlinear features and thus provide better recognition.

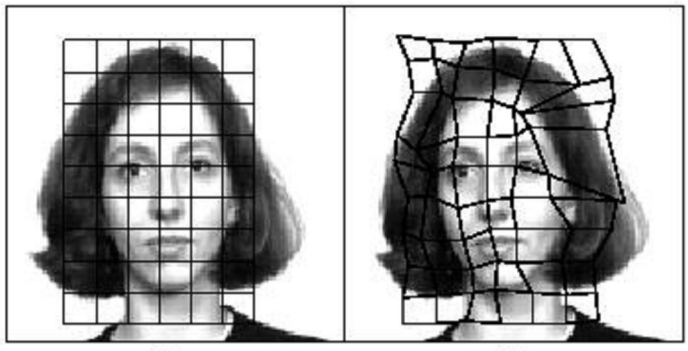
2.2.2 Model-based approaches

The model-based face recognition scheme is aimed at constructing a model of the human face, which is able to capture the facial variations. The prior knowledge of human face is highly utilized to design the model. For example, feature-based matching derives distance and relative position features from the placement of internal facial elements (e.g., eyes, etc.). Kanade [Kan73] developed one of the earliest face recognition algorithms based on automatic feature detection. By localizing the corners of the eyes, nostrils, etc. in frontal views, his system computed parameters for each face, which were compared (using a Euclidean metric) against the parameters of known faces. A more recent feature-based system, based on elastic bunch graph matching, was developed by Wiskott et al. [WFKM97] as an extension to their original graph matching system [LVBM93]. By integrating both shape and texture, Cootes et al. [LTC97] [CET01] developed a 2D morphable face model, through which the face variations are learned. A more advanced 3D morphable face model is explored to capture the true 3D structure of human face surface. The model-based scheme usually contains three steps: 1)

Constructing the model; 2) Fitting the model to the given face image; 3) Using the parameters of the fitted model as the feature vector to calculate the similarity between the query face and prototype faces in the database to perform the recognition.

2.2.2.1 Elastic graph matching

Lades et al. [LVBM93] proposed an approach for face recognition using Gabor filters called Dynamic Link Architecture (DLA). In this approach a face is represented by a labeled graph. The graph is a rectangular grid placed on the image (Figure 2.3) where nodes are labeled with responses of Gabor filters in several orientations and several spatial frequencies called *jets*. The edges are labeled with distances, where each edge connects two nodes on the graph. Comparing two faces is accomplished by adapting and matching the graph of a reference image to the graph of the test image.



(a) (b)

Figure 2-3: Example of grid matching. (a) Reference grid, (b) matched grid, [LVBM93].

Later on, Wiskott et al [WFKM97] extended DLA to Elastic Bunch Graph Matching (EBGM), where graph nodes are located at a number of selected facial landmarks, see Figure 2-4. The EBGM has shown very competitive performance and been ranked as the top method in a previous FERET evaluation [PMRR00]. The goal of Elastic graph

.

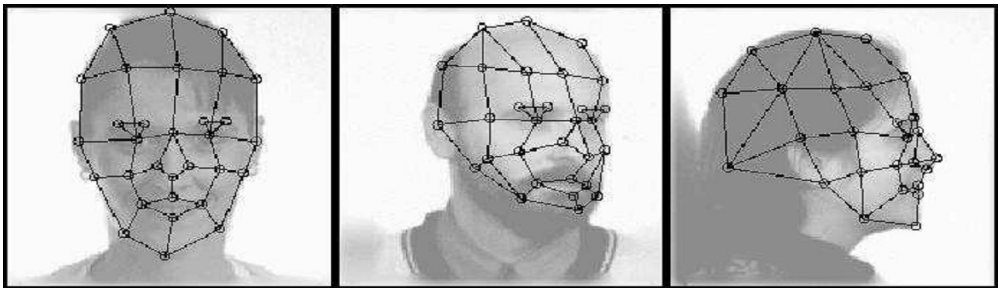


Figure 2-4: Adapted graphs for faces in different views [WFKM97].

matching is to find the fiducial points on a query image and thus to extract from the image a graph which maximizes the graph similarity function. This is performed automatically if the face bunch graph is appropriately initialized. A face bunch graph consists of a collection of individual face model graphs combined into a stack-like structure, in which each node contains the jets of all previously initialized faces from the database. To position the grid on a new face, the graph similarity between the image graph and the existing face bunch graph is maximized. Graph similarity is defined as the average of the best possible match between the new image and any face stored within the face bunch graph minus a topographical term (see Equation 2.4), which accounts for distortion between the image grid and the face bunch graphs. The similarity 'S_ϕ' between two jets is defined as:

$$S_\phi(J, J') = \frac{\sum_j a_j a'_j \cos(\phi_j - \phi'_j - \vec{dk}_j)}{\sqrt{\sum_j a_j^2 \sum_j a'^2_j}} \tag{2.3}$$

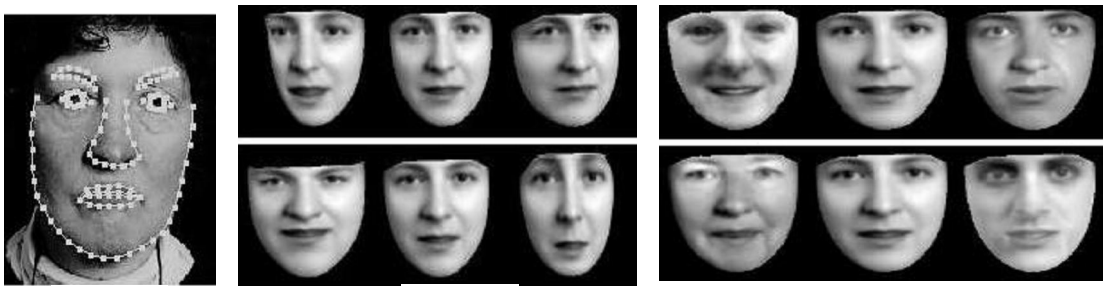
where a_j and a'_j are magnitude and phase of the Gabor coefficients in the jth jet, respectively; $\vec{d}$ is the displacement between locations of the two jets, k determines the wavelength and orientation of the Gabor wavelet kernels. For an image graph G^I with nodes n = 1,….,N and edges e = 1,……,E and a face bunch graph B with model graphs m = 1,……,M, the graph similarity is defined as:

$$S_B(G^I, B) = \frac{1}{N}\sum_n \max S_\phi(J_n^I, J_n^{Bm}) - \frac{\lambda}{E}\sum_e \frac{(\Delta\vec{x}_e^I - \Delta\vec{x}_e^B)^2}{(\Delta\vec{x}_e^B)^2} \quad (2.4)$$

where λ determines the relative importance of jets and metric structure, J_n is the jets at nodes n, and $\Delta\vec{x}_e$ is the distance vector used as labels at edges e. After the grid has been positioned on the new face, the face is identified by comparing the similarity between that face and every face stored in the bunch graph. Graphs can be easily translated, rotated, scaled, and elastically deformed, thus compensating for the variance in face images. Based on the elastic graph matching framework, a number of variations have been proposed in the literature [MH03] [DFB99] [LL00] [GIAA03] [WQ02].

2.2.2.2 Active appearance and shape model

Another prominent work is the Active Appearance Model (AAM) proposed by [LTC97]. An Active Appearance Model (AAM) is an integrated statistical model which combines a model of shape variation, active shape model (ASM) with a model of the appearance variations in a shape-normalized frame. An AAM contains a statistical model of the shape and gray-level appearance of the object of interest which can generalize to almost any valid example. Matching to an image involves finding model parameters which minimize the difference between the image and a synthesized model example, projected onto the image. The potentially large number of parameters makes this a difficult problem. To recognize a face image, both ASM and AAM are adjusted to fit the new image, which generates a number of shape and texture parameters. Those parameters, together with the local profiles at model points, are used for face recognition. The AAM is constructed based on a training set of labeled images, where landmark points are marked on each example face at key positions to outline the main features, shown in Figure 2-5, along with the landmarks the effects of varying the first two parameters of shape and appearance models are also shown.



(a) (b) (c)

Figure 2-5: (a) Landmarks for AAM, (b) variance of facial shape (c) variance of facial appearance, [LTC97].

The shape of a face is represented by a vector consisting of the positions of the landmarks, $s = (x_1,y_1,.....,x_n,y_n)^T$, where (x_i,y_i) denotes the 2D image coordinate of the i^{th} landmark point. All shape vectors of faces are normalized into a common coordinate system. The principal component analysis is applied to this set of shape vectors to construct the face shape model, denoted as: $s = \bar{s} + P_s b_s$, where s is a shape vector, $\bar{s}$ is the mean shape, Ps is a set of orthogonal modes of shape variation and b_s is a set of shape parameters. In order to construct the appearance model, the example image is warped to make the control points match the mean shape. Then the warped image region covered by the mean shape is sampled to extract the gray level intensity (texture) information. Similar to the shape model construction, a vector is generated as the representation, $g = (I_1,....,I_m)^T$ where I_i denotes the intensity of the sampled pixel in the warped image. PCA is also applied to construct a linear model, $g = \bar{g} + P_g b_g$, where $\bar{g}$ is the mean appearance vector, P_g is a set of orthogonal modes of gray-level variation and b_g is a set of gray-level model parameters. Thus, all shape and texture of any example face can be summarized by the vectors b_s and b_g. The combined model is the concatenated version of b_s and b_g, denoted as follows:

$$b = \begin{pmatrix} W_s b_s \\ b_g \end{pmatrix} = \begin{pmatrix} W_s P_s^T (s - \bar{s}) \\ P_g^T (g - \bar{g}) \end{pmatrix} \tag{2.5}$$

where, W_s is a diagonal matrix of weights for each shape parameter, allowing for the difference in units between the shape and gray scale models.

The model was built based on 300 training face images, each with 122 landmark points [ECT98]. A shape model with 23 parameters, a shape-normalized texture model with 113 parameters and a combined appearance model with 80 parameters are generated.

For all the training images, the corresponding model parameter vectors are used as the feature vectors. The linear discrimination analysis is utilized to construct the discriminant subspace for face identity recognition. Given a query image, the AAM fitting is applied to extract the corresponding feature vector. The recognition is achieved by finding the best match between the query feature vector and stored prototype feature vectors, both of which are projected onto the discriminant subspace [CET01]. An example of the AAM fitting is shown in Figure 2-6.

Figure 2-6: Example of AAM fitting, [CET01]

2.2.2.3 Other statistical models

In the 2D context, among others, generative models based on Hidden Markov Modeling 'HMM' and Gaussian mixtures models 'GMM' have also been employed. HMM are widely used to learn the state and transitional probabilities between a number of hidden states. HMMs are normally trained from examples that are represented by a sequence of observations. The parameters of the HMM are initialized and then adjusted to maximize the probability of the observation of the given training samples. Samaria and Young [SY94] first proposed a HMM architecture for face recognition. A face pattern is divided into several regions such as forehead, eyes, nose, mouth and chin. These regions occur in the natural order from top to bottom and they are used to form the hidden states of 1D or pseudo 2D HMMs. To train a HMM, each face image is represented by a sequence of observation vectors, which are constructed from the pixels of a sub window. Nefian and Hayes [NH99] proposed the embedded 2D HMM, which consists of a set of super states with each super state being associated with a set of embedded states. Super states represent primary facial regions while embedded states within each super state describe in more detail the facial regions. However, HMM based systems require lots of images for training, and are only capable of operating on small databases. The performance drops dramatically as the size of database is scaled up. For example, performance of Nefian and Hayes's method drops from 97.5% to 32.5% when the number of subjects rises from 40 to 200 [BS03]. Another important issue associated with HMM based approaches is the model training process. Usually when there is a database update, e.g. there are new faces or images to be enrolled or to be removed, the HMMs need to be retrained. This issue affects the flexibility of HMM systems with large or frequently updated database. Compared to 1D and pseudo 2D HMM, very recently [Hun08] proposed an extension called Joint Multiple-HMM. This approach offers computational advantages and the good learning ability from just a single sample per class. The new method also includes some improvements over its previous counterparts and does not require retraining HMMs for new images or subjects.

In another recent work, Gaussian mixture models are used, to address the pose invariant face recognition. A generic GMM face model is first derived from an

independent database of faces for each view, and then this model is adapted for each client (person) to be recognized later. A composite model for all views is then generated for each person. The method is tested in a verification scenario and provides reasonable accuracy in near frontal views [CSB06].

2.2.3 3D face recognition

Human face is a surface lying in the 3D space intrinsically. Therefore, in principle, the 3D model is better for representing faces, especially to handle facial variations, such as pose, illumination. Blanz et al. [BV99] [BRV02] proposed a method based on a 3D morphable face model that encodes shape and texture in terms of model parameters, and an algorithm that recovers these parameters from a single image of a face. To handle the extreme image variations induced by these parameters, one common approach taken by various authors is to use generative image models. For image analysis, the general strategy of all these techniques is to fit the generative model to a new image, thereby parameterize it in terms of the model. In order to make identification independent of imaging conditions, the goal is to separate intrinsic model parameters of the face from extrinsic imaging parameters. The separation of intrinsic and extrinsic parameters is achieved explicitly by simulating the process of image formation using 3D computer graphics technology. In these works, a 3D face model was usually used to synthesize images with different illumination and poses from a frontal face image, 2D techniques are then applied to the synthesized images for recognition. Figure 2-7 illustrates the scheme.

Similar work can also be found in [ZC00] [LR03]. With the development of 3D capture systems, face recognition using 3D facial data is also attracting much attention. [BA00] developed both surface matching and central/lateral profiles for recognition, the results show that the two methods give the same level of performance. Some works also applied 2D techniques to 3D range data for recognition, e.g., 3D Eigenfaces [HE02]. In addition to using 3D data only, multi-modal 3D+2D face recognition has also been proposed [WCH02].

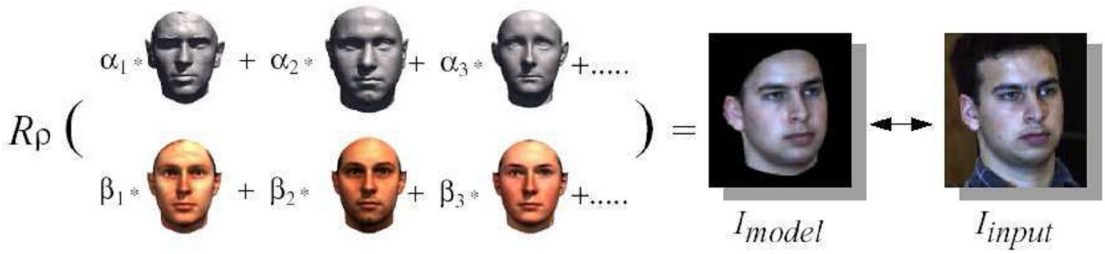


Figure 2-7: The goal of the 3D model fitting process is to find shape and texture coefficients α and β such that rendering R_p produces an image I_{model} that is as similar as possible to I_{input} [BRV02].

While 3D shape is defined independent of illumination, it is sensed dependent of illuminations. "Holes" may occur in areas where data is missing [BCF06], even under ideal illuminations, 3D depth resolution also needs to be improved to benefit the recognition algorithms.

2.2.4 Other approaches

A large number of other important approaches for face recognition systems have been investigated, here we have reviewed the most important and key ideas. Some more recent methods addressing view-point invariant face recognition will be discussed in part 3 of this thesis. For a more complete list on the existing literature, see some of the recent face recognition literature reviews [ZCRP03] [ANRS07].

2.3 Facial Databases and Protocols

A number of face databases have been collected for different face recognition tasks and the choice of the ones used in order to evaluate the performances has to be considered carefully. Actually, there is no ideal database and the test data should represent as loosely as possible the data encountered in real life and thus depends on the final application. A large variability occurs in the existing face databases in terms of camera quality, time lapse between the different images and capture of images under variable environment. The variable environment includes head pose variability, illumination conditions, background, expression etc.

Keeping in mind these parameters, we have chosen two large and comprehensive face databases, FERET and CMU-PIE face databases that are publically available and includes thousands of images depicting large variations due to pose and illumination and expression. The choice is also motivated by the fact that most of the competitive face recognition methods use these to report the performance, and this provides a standard benchmark to compare our method with that of others. Despite these, another face database, the ORL face database, is also used. The ORL face database however is a small database (in terms of the variability presented in images of each person) and is primarily used for preliminary experiments and some validation and parameter setting procedures. However, all these three databases are very popular and being used by the face recognition community in order to design, evaluate and compare their systems.

With the primary focus of this thesis to develop techniques for recognition across pose and illumination variations, the chosen databases provide a very good platform.

2.3.1 The ORL database

The Olivetti Research Ltd. (ORL) database (http://www.cam-orl.co.uk) contains images from 40 individuals, each providing 10 different images. Each image contains a face area mainly (with little background). All the images are grayscale at a resolution of 92 x 112 pixels. The 10 different images of each person include variations of pose (some tilting and rotation of the face of up to 20 degrees) and facial expressions, while the illumination is almost constant. In addition, there is variation in the scale of up to about 10 percent. Moreover, the images are not aligned. About half of the faces are upright and have a small rotation on the y-axis. The other half of the faces show different amounts of perspective variations. Thus the ORL face images are very similar to those produced from an automatic face detection system in terms of precision in localization, scale, pose and normalization. Figure 2-8 shows all 10 examples per subject for a number of subjects.

Figure 2-8: Example images of different subjects in ORL face database

2.3.2 The CMU-PIE database

The CMU Pose Illumination Expression (PIE) database [SBB02] contains a total of 41,368 images taken from 68 individuals. The subjects were imaged in the CMU 3D Room using a set of 13 synchronized high-quality color cameras and 21 flashes. The resulting images are 640x480 in size, with 24-bit color resolution. Each of the 13 poses is separated by approximately 22.5^{o}, and thus varies from full left profile to frontal and so on to full right profile. Some typical pose variations for one subject are shown in figure 2-9. Four expression variations are captured in each pose for each subject. These are neutral, smiling, blinking and talking. Illumination variations are captured for each subject in each pose with each of 21 flashes both with and without ambient lighting, see figure 2-10 for an example. 4 of the 13 cameras (c25,c09, c07,and c37, figure 2-9) are placed above the head in frontal and ¾ profile views, so as to capture the up and down tilt of the face. This is done to mimic a typical surveillance camera capture situation.

Figure 2-9: An example of pose variation in the CMU PIE database. Images of one person from 8 of the 13 cameras are shown. The other 5 camera are arranged symmetrically to the 5 cameras on the left. The pose varies from full left profile to full frontal and so on to full right profile [SBB02].

Figure 2-10: PIE expression and Illumination variations captured in each pose for each person. (a) Different expression variations (b) Different illumination conditions with ambient lighting (c) Different illumination conditions without ambient lighting.

2.3.3 The FERET database

The Facial Recognition Technology database 'FERET' [PMRR00] has been distributed primarily for documenting the advances in automatic face recognition and is administered by National Institute of Standards and Technology 'NIST' in United States. This is the most comprehensive publically available database and contains a large number of subjects captured under different expression, pose, illumination and aging. The most common FERET protocol defines evaluation strategy by giving standard training and test sets, mainly used in frontal face recognition scenarios in different illumination, expression and aging effects. While, here we used the subset of FERET database that concerns pose variations, called the pose subset. The database contains a total of 14,126 images that includes 1199 individuals. Where, the pose subset includes images of 200 persons in 9 different poses. Tabel 2-1 summarize the distribution and nomenclature of the pose subset used in this thesis. Figure 2-11 shows typical pose variations for one of the subjects in FERET.

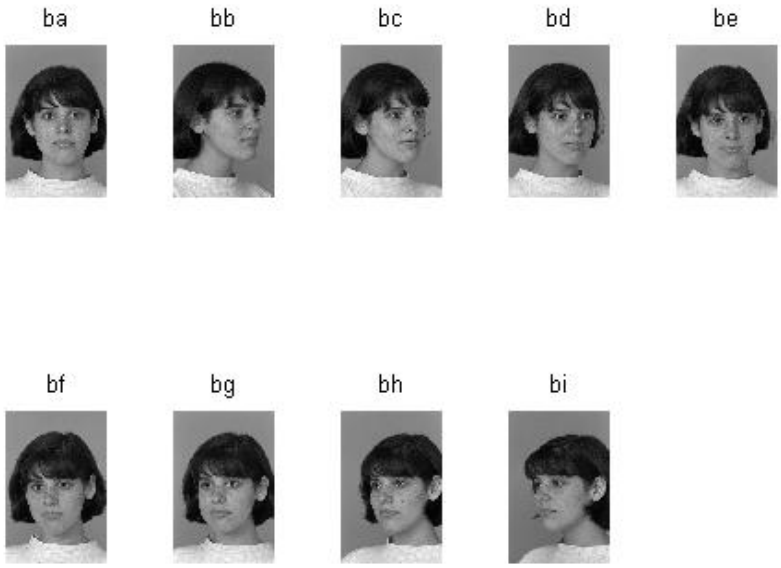


Figure 2-11: Pose variations in FERET pose subset.

Table 2-1: Nomenclature of Pose subset in FERET Database [PMRR00]

Two letter code	*Pose Angle (degrees)*	*Description*	*Number in Database*	*Number of Subjects*
fa	0 = frontal	Regular facial expression	1762	1010
fb	0	Alternative facial expression	1518	1009
ba	0	Frontal "b" series	200	200
bj	0	Alternative expression to ba	200	200
bk	0	Different illumination to ba	200	200
bb	+60	Subject faces to his left which is the photographer's right	200	200
bc	+45		200	200
bd	+25		200	200
be	+15		200	200
bf	-15	Subject faces to his right which is the photographer's left	200	200
bg	-25		200	200
bh	-45		200	200
bi	-60		200	200
ql	-22.5	Quarter left and right	763	508
qr	+22.5		763	508
hl	-67.5	Half left and right	1246	904
hr	+67.5		1298	939
pl	-90	Profile left and right	1318	974
pr	+90		1342	980

2.4 Conclusion

In this chapter we have elaborated on the background and specifics of an automatic face recognition process. More specifically, we differentiate the face localization from alignment and argue that the alignment stage has to be avoided in order to realize a fully automatic system especially when considering the pose variations. A brief but comprehensive literature review of the current state of the art methods is presented, and the facial databases used along this thesis have been described.

Chapter 3. Feature Extraction for Face Recognition

The goal of feature extraction is to find a specific representation of the data that can highlight relevant information. This representation can be found by maximizing a criterion or can be a pre-defined representation. Usually, a face image is represented by a high dimensional vector containing pixel values (holistic representation) or a set of vectors where each vector summarizes the underlying content of a local region by using a high level transformation (local representation). Typically, the vectors are projected into a new space (the feature space), then, the least relevant features can be removed to reduce the dimension of the feature vector according to some criterion.

In this chapter we present an overview of the most relevant feature extraction techniques used in face recognition and propose a feature extraction technique, to be used in face recognition for the first time, which has a number of advantages over the commonly used feature descriptions in the context of face recognition.

3.1 Holistic Vs Local Features-What Features to Use?

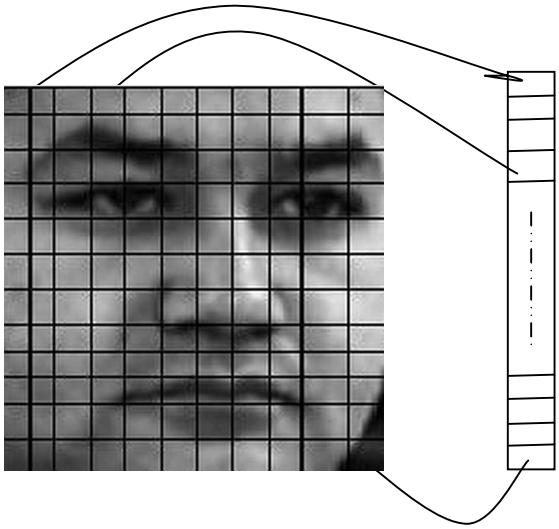
Holistic representation is the most typical to be used in face recognition. It is based on lexicographic ordering of raw pixel values to yield one vector per image. An image can now be seen as a point in a high dimensional feature space. The dimensionality corresponds directly to the size of the image in terms of pixels. Therefore, an image of size 100x100 pixels can be seen as a point in a 10,000 dimensional feature space. This large dimensionality of the problem prohibits the use of any learning to be carried out in such a high dimensional feature space. This is called the curse of dimensionality in the pattern recognition literature [DHS01]. A common way of dealing with it is to employ a dimensionality reduction technique such as PCA to pose the problem into a low-

dimensional feature space such that the major modes of variation of the data are still preserved.

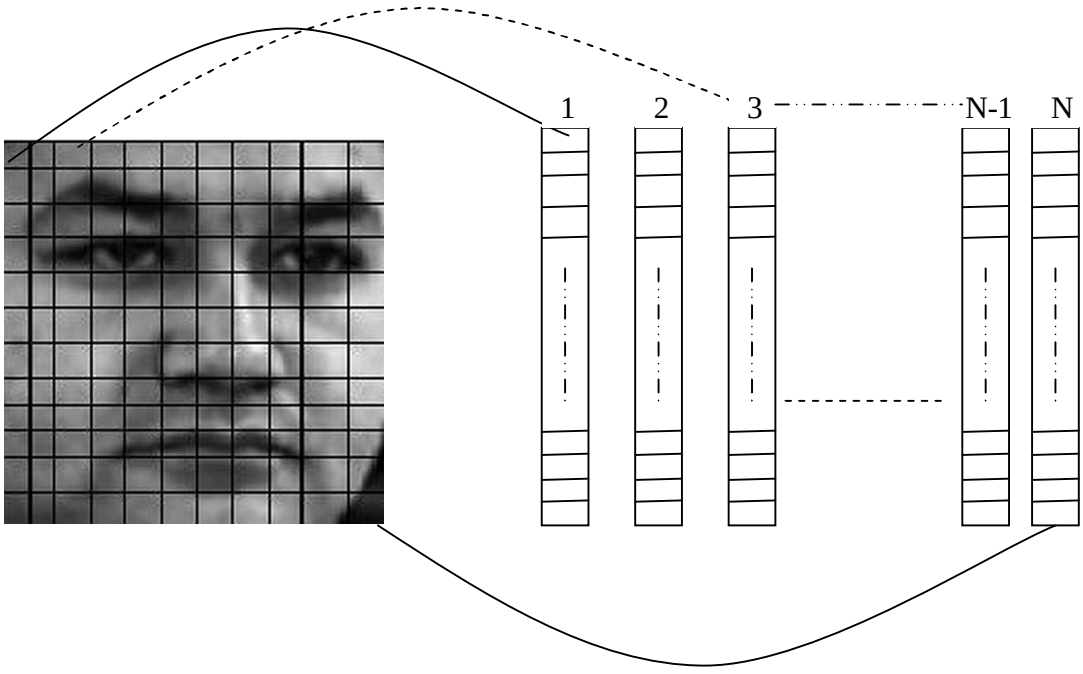
Local feature extraction refers to describing only a local region/part of the image by using some transformation rule or specific measurements such that the final result describes the underlying image content in a manner that should yield a unique solution whenever the same content is encountered. In doing so, however it is also required to have some degree of invariance with respect to commonly encountered variations such as translation, scale and rotations.

A number of authors [PMS94] [CSB06] [ZJN07] do not differentiate the holistic and local approaches according to the very nature they are obtained, but rather use the terms in lieu of global (having one feature vector per image) and a bag-of-feature (having several feature vectors per image) respectively. Here we want to put the both terms into their right context, and hence a holistic representation can be obtained for several local regions of the image and similarly a local representation can still be obtained by concatenating several locally processed regions of the image into one global vector, see figure 3-1 for an illustration. An example of the first usage is local-PCA or modular-PCA [GA04] [TC05], where an image is divided into several parts or regions, and each region is then described by a vector comprising underlying raw-pixel values, PCA is then employed to reduce the dimensionality. Note that it is called local since it uses several local patches of the same image but it is still holistic in nature. An example of the second is what usually found in the literature, e.g. Gabor filtering, Discrete Cosine Transform 'DCT', Local Binary Pattern 'LBP' etc where each pixel or local region of the image is described by a vector and concatenated into a global description [ZJN07], note that they still give rise to one vector per image but they are called local in the literature because they summarize the local content of the image at a location in a way that is invariant with respect to some intrinsic image properties e.g. scale, translation and/or rotation.

Keeping in view the above discussion it is common in face recognition to either follow a global feature extraction or a bag-of-features approach. The choice, of what is optimal, depends on the final application in mind and hence is not trivial. However, there are a number of advantages and disadvantages with both the approaches.



One global vector per image obtained by concatenating pixels (holistic) or processed local regions/patches (local).



A “bag-of-features” approach, where N vectors are obtained for N local patches/regions. Each feature vector may be obtained by holistic or local feature extraction.

Figure 3-1: Global and bag-of-feature representation for a face image.

For instance, a global description is generally preferred for face recognition since it preserves the configural (i.e., the interrelations between facial parts) information of the face, which is very important for preserving the identity of the individual as have been evidenced both from psychological [MCVC06], neurobiological [SWCG06] [HRS08] and computer vision [BHK97] [CLLH01] communities. On the other hand, a bag-of-features approach has been taken by a number of authors [BP93] [M02] [KY03] and shown improved recognition results in the presence of occlusion etc., nonetheless in doing so, these approaches are bound to preserve the configural information of the facial parts either implicitly or explicitly by comparing only the corresponding parts in two images and hence put a hard demand on the requirement of proper and precise alignment of facial images.

Note that while occlusion may be the one strong reason to consider a bag-of-features approach, the tendency of preserving the spatial arrangement of different facial parts (configural information) is largely compromised. As evidenced from the many studies from interdisciplinary fields that this spatial arrangement is in fact quite crucial in order to preserve the identity of an individual, we therefore, advocate the use of a global representation for a face image in this dissertation, as has also been used by many others.

One may, however, note that a global representation does not necessarily mean a holistic representation, as described before. In fact, for the automatic unconstrained face recognition, where there may be much variation in terms of scale, lighting, misalignments etc, the choice of using local feature extraction becomes imperative since holistic representation can not generalize in these scenarios and is known to be highly affected by these in-class variations.

3.2 Holistic Feature Extraction

Holistic feature extraction is the most widely used feature description technique in appearance based face recognition methods. Despite its poor generalization abilities in unconstrained scenarios, it is being used for the main reason that any local extraction

technique is a form of information reduction in that it typically finds a transformation that describes a large data by few numbers. Since from a strict general object recognition stand point, face is one class of object, and thus discriminating within this class puts very high demands in finding subtle details of an image that discriminates among different faces. Therefore each pixel of an image is considered valuable information and holistic processing develops. However, a holistic-based global representation as been used classically [TP91] can not perform well and therefore more recently many researchers used a bag-of-features approach, where each block or image patch is described by holistic representation and the deformation of each patch is modeled for each face class [KY03] [LC06] [ALC08].

3.2.1 Eigenface-A global representation

Given a face image matrix *F* of size *Y* x *X*, a vector representation is constructed by concatenating all the columns of *F* to form a column vector $\vec{f}$ of dimensionality *YX*. Given a set of training vectors $\{\vec{f}_i\}_{i=1}^{Np}$ for all persons, a new set of mean subtracted vectors is formed using:

$$g_i = \vec{f}_i - \vec{f}_\mu, \qquad i = 1, 2, ..., N_p \tag{3.1}$$

The mean subtracted training set is represented as a matrix $G = [\vec{g}_1, \vec{g}_2, ..., \vec{g}_{Np}]$. The covariance matrix is then calculated using, $\sum = GG^T$. Due to the size of $\sum$, calculation of the eigenvectors of $\sum$ can be computationally infeasible. However, if the number of training vectors (N_P) is less than their dimensionality (*YX*), there will be only N_p-1 meaningful eigenvectors. Turk and Pentland [TP91] exploit this fact to determine the eigenvectors using an alternative method summarized as follows. Let us denote the eigenvectors of matrix G^TG as $\vec{v}_j$ with corresponding eigenvalues Λ_j:

$$G^T G\vec{v}_j = \Lambda_j \vec{v}_j \tag{3.2}$$

Pre-multiplying both sides by *G* gives us: $GG^T G\vec{v}_j = \Lambda_j G\vec{v}_j$, Letting $\vec{e}_j = G\vec{v}$ and substituting for $\sum$ from equation 3.1:

$$\sum \vec{e}_j = \Lambda_j \vec{e}_j \tag{3.3}$$

Hence the eigenvectors of $\sum$ can be found by pre-multiplying the eigenvectors of G^TG by *G*. To achieve dimensionality reduction, let us construct matrix $E = [\vec{e}_1, \vec{e}_1, ..., \vec{e}_D]$, containing D eigenvectors of $\sum$ with largest corresponding eigenvalues. Here, D<N_p, a feature vector $\vec{x}$ of dimensionality D *is* then derived from a face vector $\vec{f}$ using:

$$\vec{x} = E^T(\vec{f} - \vec{f}_\mu) \tag{3.4}$$

Therefore, a face vector $\vec{f}$ is decomposed into D eigenvectors, known as eigenfaces.

Similarly, employing the above mentioned Eigen analysis to each local patch of the image results into a bag-of-features approach. Pentland *et al.* extended the eigenface technique to a layered representation by combining eigenfaces and other eigenmodules, such as eigeneyes, eigennoses, and eigenmouths [PMS94]. Recognition is then performed by finding a projection of the test image patch to each of the learned local Eigen subspaces for every individual.

3.3 Local Feature Extraction

[GA04] argued that some of the local facial features did not vary with pose, direction of lighting and facial expression and, therefore, suggested dividing the face region into smaller sub images. The goal of local feature extraction thus becomes to represent these local regions effectively and comprehensively. Here we present the most commonly used local feature extractions techniques in face recognition namely the Gabor wavelet transform based features , discrete cosine transform DCT-based features and more recently proposed Local binary pattern LBP features.

3.3.1 2D Gabor wavelets

The 2D Gabor elementary function was first introduced by Granlund [Gra78]. Gabor wavelets demonstrate two desirable characteristic: spatial locality and orientation selectivity. The structure and functions of Gabor kernels are similar to the two-dimensional receptive fields of the mammalian cortical simple cells [HW78]. [OF96]

[RB95] [SC00] indicates that the Gabor wavelet representation of face images should be robust to variations due to illumination and facial expression changes. Two-dimensional Gabor wavelets were first introduced into biometric research by Daugman [D93] for human iris recognition. Lades et al. [LVBM93] first apply Gabor wavelets for face recognition using the Dynamic Link Architecture framework.

A Gabor wavelet kernel can be thought of a product of a complex sinusoid plane wave with a Gaussian envelop. A Gabor wavelet generally used in face recognition is defined as [Liu04]:

$$\psi_{u,v}(z) = \frac{\left\| k_{u,v} \right\|^2}{\sigma^2} e^{-\frac{\left\| k_{u,v} \right\|^2 \left\| z \right\|^2}{2\sigma^2}} [e^{ik_{u,v}z} - e^{-\frac{\sigma^2}{2}}] \tag{3.5}$$

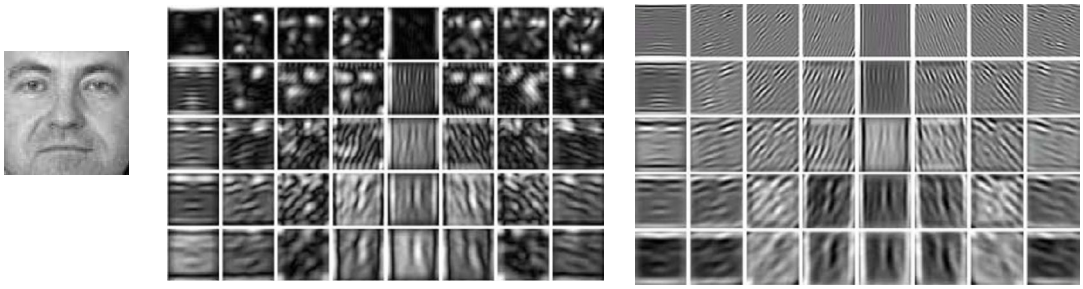
where z = (x, y) is the point with the horizontal coordinate x and the vertical coordinate y in the image plane. The parameters u and v define the orientation and frequency of the Gabor kernel, $\|.\|$ denotes the norm operator, and σ is related to the standard derivation of the Gaussian window in the kernel and determines the ratio of the Gaussian window width to the wavelength. The wave vector $k_{\mu,v}$ is defined as $k_{u,v} = k_v e^{i\phi_u}$.

Following the parameters suggested in [LVBM93] and used widely in prior works [Liu04] [LW02], $k_v = \frac{k_{\max}}{f^v}$ and $\phi_u = \frac{\pi u}{8}$. k_{max} is the maximum frequency, and f^v is the spatial frequency between kernels in the frequency domain. $v \in \{0,...,4\}$ and $u \in \{0,...,7\}$ in order to have a Gabor kernel tuned to 5 scales and 8 orientations. Gabor wavelets are chosen relative to $\sigma = 2\pi$, $k_{\max} = \frac{\pi}{2}$ and $f = \sqrt{2}$. The parameters ensures that frequencies are spaced in octave steps from 0 to π, typically each Gabor wavelet has a frequency bandwidth of one octave that is sufficient to have less overlap and cover the whole spectrum.

The Gabor wavelet representation of an image is the convolution of the image with a family of Gabor kernels as defined by equation (3.5). The convolution of image *I* and a Gabor kernel $\psi_{u,v}(z)$ is defined as follows:

$$G_{u,v}(z) = I(z) * \psi_{u,v}(z) \tag{3.5}$$

where $z = (x, y)$ denotes the image position, the symbol '$*$' denotes the convolution operator, and $G_{u,v}(z)$ is the convolution result corresponding to the Gabor kernel at scale v and orientation u . The Gabor wavelet coefficient is a complex with a real and imaginary part, which can be rewritten as $G_{u,v}(z) = A_{u,v}(z).e^{i\theta_{u,v}(z)}$, where $A_{u,v}$ is the magnitude response and $\theta_{u,v}$ is the phase of Gabor kernel at each image position. It is known that the magnitude varies slowly with the spatial position, while the phases rotate in some rate with positions, as can be seen from the example in figure 3-2. Due to this rotation, the phases taken from image points only a few pixels apart have very different values, although representing almost the same local feature [WFKM97]. This can cause severe problems for face matching, and it is just the reason that all most all of the previous works make use of only the magnitude part for face recognition.



(a) (b)

Figure 3-2: Visualization of Gabor magnitude (a) and phase response (b) for a face image with 40 Gabor wavelets (5 scales and 8 orientations).

Note that, convolving an image with a bank of Gabor kernel tuned to 5 scales and 8 orientations results in 40 magnitude and phase response maps of the same size as image. Therefore, considering only the magnitude response for the purpose of feature description, each pixel can be now described by a 40 dimensional feature vector (by concatenating all the response values at each scale and orientation) describing the response of Gabor filtering at that location.

Note that Gabor feature extraction results in a highly localized and over complete response at each image location. In order to describe a whole face image by Gabor feature description the earlier methods take into account the response only at certain

image locations, e.g. by placing a coarse rectangular grid over the image and taking the response only at the nodes of the grid [LVBM93] or just considering the points at important facial landmarks as in Wiskot et al [WFKM97]. The recognition is then performed by directly comparing the corresponding points in two images. This is done for the main reason of putting an upper limit on the dimensionality of the problem. However, in doing so they implicitly assume a perfect alignment between all the facial images, and moreover the selected points that needs to be compared have to be detected with pixel accuracy.

One way of relaxing the constraint of detecting landmarks with pixel accuracy is to describe the image by a global feature vector either by concatenating all the pixel responses into one long vector or employ a feature selection mechanism to only include significant points [WYS02] [LLS04] . One global vector per image results in a very high and prohibitive dimensional problem, since e.g. a 100x100 image would result in a 40x100x100=400000 dimensional feature vector. Some authors used Kernel PCA to reduce this dimensionality termed as Gabor-KPCA [Liu04], and others [WYS02] [LLS04] [WCH02] employ a feature selection mechanism for selecting only the important points by using some automated methods such as Adaboost etc.. Nonetheless, a global description in this case still results in a very high dimensional feature vector, e.g. in [WCH02] authors selected only 32 points in an image of size 64x64, which results in 32x40=1280 dimensional vector, due to this high dimensionality the recognition is usually performed by computing directly a distance measure or similarity metric between two images. The other way can be of taking a bag-of-feature approach where each selected point is considered an independent feature, but in this case the configural information of the face is effectively lost and as such it can not be applied directly in situations where a large pose variations and other appearance variations are expected.

The Gabor based feature description of faces although have shown superior results in terms of recognition, however we note that this is only the case when frontal or near frontal facial images are considered. Due to the problems associated with the large dimensionality, and thus the requirement of feature selection, it can not be applied directly in scenarios where large pose variations are present.

3.3.2 2D Discrete cosine transform

Another popular feature extraction technique has been to decompose the image on block by block basis and describe each block by 2D Discrete Cosine Transform 'DCT' coefficients. An image block $f(p,q)$, where $p,q = \{0,1..,N-1\}$ (typically N=8), is decomposed terms of orthogonal 2D DCT basis functions. The result is a NxN matrix $C(v,u)$ containing 2D DCT coefficients:

$$C(v,u) = \alpha(v)\alpha(u)\sum_{y=0}^{N-1}\sum_{x=0}^{N-1} f(p,q)\beta(p,q,v,u) \tag{3.6}$$

where $v,u = 0,1,2,...,N-1$, $\alpha(v) = \sqrt{\frac{1}{N}}$ for v=0, and $\alpha(v) = \sqrt{\frac{2}{N}}$ for v=1,2,…,N-1 and

$$\beta(p,q,v,u) = \cos\left[\frac{(2p+1)v\pi}{2N}\right]\cos\left[\frac{(2q+1)u\pi}{2N}\right] \tag{3.7}$$

The coefficients are ordered according to a zig-zag pattern, reflecting the amount of information stored [GW93]. For a block located at image position (x,y), the baseline 2D DCT feature vector is composed of:

$$x = [c_o^{(x,y)} \quad c_1^{(x,y)} ... \quad c_{M-1}^{(x,y)}]^T \tag{3.8}$$

Where $c_n^{(x,y)}$ denotes the *n*-th 2D DCT coefficient and *M* is the number of retained coefficients[3]. To ensure adequate representation of the image, each block overlaps its horizontally and vertically neighboring blocks by 50% [EMR00]. M is typically set to 15 therefore each block yields a 15 dimensional feature vector. Thus for an image which has Y rows and X columns, there are $N_D = (2\frac{Y}{N}-1)\times(2\frac{X}{N}-1)$ blocks [sanderson04].

DCT based features have mainly been used in HMM based methods in frontal scenarios. More recently [] proposed an extension of conventional DCT based features by replacing the first 3 coefficients with there corresponding horizontal and vertical deltas termed as DCTmod2, resulting into an 18-dimensional feature vector for each block. The authors claimed that this way the feature vectors are less affected by illumination change. They then use a bag-of-feature approach to derive person specific face models by using GMM.

3.3.3 Local Binary Pattern Histogram and other features

Local binary pattern (LBP) was originally designed for texture classification [OPM02] , and was introduced in face recognition in [AHP04]. As mentioned in [AHP04] the operator labels the pixels of an image by thresholding some neighborhood of each pixel with the center value and considering the result as a binary number. Then the histogram of the labels can be used as a texture descriptor. See figure 3-3 for an illustration of the basic $LBP_{P,R}^{U2}$ operator. The face area is divided into several small windows. Several LBP operators are compared and $LBP_{8,2}^{U2}$ the operator in 18x21 pixel windows is recommended because it is a good tradeoff between recognition performance and feature vector length. The subscript represents using the operator in a (P, R) neighborhood. Superscript U2 stands for using only uniform patterns and labeling all remaining patterns with a single label, see [AHP04] for details. The chi square statistic and the weighted chi square statistic were adopted to compare local binary pattern histograms.

Recently Zhang *et al.* [ZSGCZ05] proposed local Gabor binary pattern histogram sequence (LGBPHS) by combining Gabor filters and the local binary operator. [BSXW07] further used LBP to encode Gabor filter phase response into an image histogram termed as Histogram of Gabor Phase Patterns (HGPP).

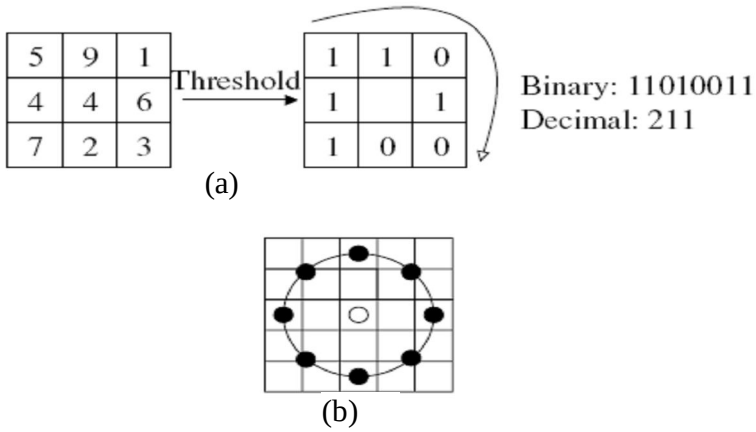


Figure 3-3: (a) the basic LBP operator. (b) The circular (8,2) neighborhood. The pixel values are bilinearly interpolated whenever the sampling point is not in the center of a pixel [AHP04].

3.4 Face-GLOH-Signatures –Introduced feature representation for face recognition

The mostly used local feature extraction and representation schemes presented in previous section have mainly been employed in a frontal face recognition task. There ability to perform equally well when a significant pose variation is present among images of the same person can not be guaranteed, especially when no alignment is assumed among facial images. This is because when these feature representations are used as a global description the necessity of having a precise alignment becomes unavoidable. While representations like 2D-DCT or LBP are much more susceptible to noise, e.g. due to illumination change as noted in Jie Zou et al [ZJN07] or pose variations, Gabor based features are considered to be more invariant with respect to these variations. However, as discussed earlier the global Gabor representation results in a prohibitively high dimensional problem and as such can not be directly used in statistical based methods to model these in-class variations due to pose for instance. Moreover the effect of misalignments on Gabor features has been studied [SGCCY04], where strong performance degradation is observed for different face recognition systems.

As to the question, what description to use, there are some guidelines one can benefit from. For example, as discussed in section 3.1 the configural relationship of the face has to be preserved. Therefore a global representation as opposed to a bag-of-features approach should be preferred. Further in order to account for the in-class variations the local regions of the image should be processed in a scale, rotation and translation invariant manner. Another important consideration should be with respect to the size of the local region used. Some recent studies [M02] [UVS02] [ZLG05] show that large areas should be preferred in order to preserve the identity in face identification scenarios.

Keeping in view the preceding discussion we use features proposed in [MS05], used in other object recognition tasks, and introduce to employ these for the task of face recognition for the first time [SJH06][SH08b]. Our approach is to extract whole appearance of the face in a manner which is robust against misalignments. For this the

feature description is specifically adapted for the purpose of face recognition. It models the local parts of the face and combines them into a global description

We use a representation based on gradient location-orientation histogram (GLOH) [MS05], which is more sophisticated and is specifically designed to reduce in-class variance by providing some degree of invariance to the aforementioned transformations. GLOH features are an extension to the descriptors used in the scale invariant feature transform (SIFT) [L04], and have been reported to outperform other types of descriptors in object recognition tasks [MS05]. Like SIFT the GLOH descriptor is a 3D histogram of gradient location and orientation, where location is quantized into a log-polar location grid and the gradient angle is quantized into eight orientations. Each orientation plane represents the gradient magnitude corresponding to a given orientation. To obtain illumination invariance, the descriptor is normalized by the square root of the sum of squared components.

Originally [MS05] used the log-polar location grid with three bins in radial direction (the radius set to 6, 11, and 15) and 8 in angular direction, which results in 17 location bins. The gradient orientations are quantized in 16 bins. This gives a 272 bin histogram. The size of this descriptor is reduced with PCA.

While here the extraction procedure has been specifically adapted to the task of face recognition and is described in the remainder of this section.

The extraction process begins with the computation of scale adaptive spatial gradients for a given image *I(x,y)*. These gradients are given by:

$$\nabla_{xy} \equiv \sum_{t} w(x,y,t)\sqrt{t}\nabla^{t}_{xy} L(x,y;t) \tag{3.9}$$

where *L(x,y; t)* denotes the linear Gaussian scale space of *I(x,y)* [L94] and *w(x,y,t)* is a weighting, as given in equation 3.10.

$$w(x,y,t) = \frac{\left|\sqrt{t}\nabla^{t}_{xy} L(x,y;t)\right|^{4}}{\sum_{t}\left|\sqrt{t}\nabla^{t}_{xy} L(x,y;t)\right|^{4}} \tag{3.10}$$

The gradient magnitudes obtained for an example face image (Figure 3-4 e) are shown in Figure 3-4 b. The gradient image is then partitioned on a grid in polar coordinates, as illustrated in Figure 3-4 c. As opposed to the original descriptor the partitions include a central region and seven radial sectors. The radius of the central region is chosen to make the areas of all partitions equal. Each partition is then processed to yield a histogram of gradient magnitude over gradient orientations. The histogram for each partition has 16 bins corresponding to orientations between 0 and 2π, and all histograms are concatenated to give the final 128 dimensional feature vector, that we term as face-GLOH-signature, see Figure 3-4 d. No PCA is performed in order to reduce the dimensionality.

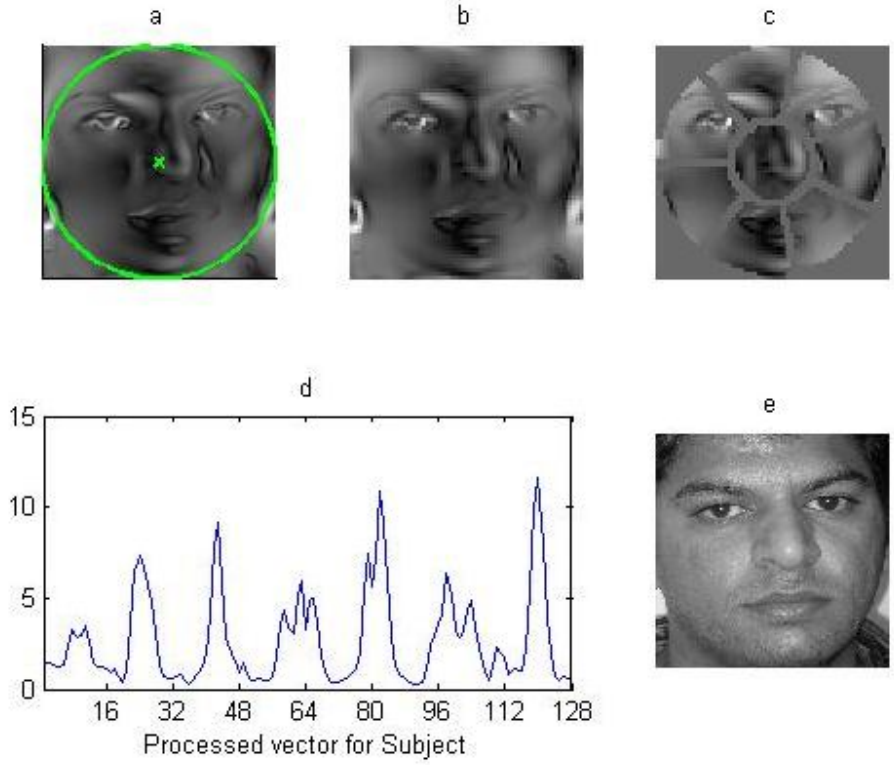


Figure 3-4: : Face-GLOH-Signature extraction (a-b) Gradient magnitudes (c) polar-grid partitions (d) 128-dimentional feature vector (e) Example image of a subject.

The dimensionality of the feature vector depends on the number of partitions used. A higher number of partitions results in a longer vector and vice versa. The choice has to be made with respect to some experimental evidence and the effect on the recognition performance. We have assessed the recognition performance on a validation set by using ORL face database. By varying the partitions sizes from 3 (1 central region and 2 sectors), 5, 8, 12 and 17, we found that increasing number of partitions results in degrading performance especially with respect to misalignments while using coarse partitions also affects recognition performance with more pose variations. Based on the results, 8 partitions seem to be the optimal choice and a good trade off between achieving better recognition performance and minimizing the effect of misalignment. The efficacy of the descriptor is demonstrated in an extensive experimental setup in the presence of pose variations and misalignments, in the next chapter.

It should be noted that, in practice, the quality of the descriptor improves when care is taken to minimize aliasing artifacts. The recommended measures include the use of smooth partition boundaries as well as a soft assignment of gradient vectors to orientation histogram bins.

3.5 Conclusion

A comprehensive account of almost all the feature extraction methods used in current face recognition systems is presented. Specifically we have made distinction in the holistic and local feature extraction and differentiate them qualitatively as opposed to quantitatively. It is argued that a global feature representation should be preferred over a bag-of-feature approach. The problems in current feature extraction techniques and there reliance on a strict alignment is discussed. Finally we have introduced to use face-GLOH signatures that are invariant with respect to scale, translation and rotation and therefore do not require properly aligned images. The resulting dimensionality of the vector is also low as compared to other commonly used local features such as Gabor, LBP etc. and therefore statistical based methods can also benefit from it. In the next chapter we will study the performance of this feature description in a typical face recognition scenario.

Chapter 4. Performance Analysis of Classifiers in Multi-view Face Recognition Scenarios

In this chapter we present a rigorous analysis of several conventional classifiers in a typical multi-view face recognition task. The objective of this chapter is to assess the performance with respect to small sample size problem as well as the effects when a proper alignment is not assumed and there exist large pose differences. The effectiveness of the face-GLOH signatures, as opposed to most commonly used feature description as introduced in chapter 3, is demonstrated in an extensive experimental setting under these more practical conditions. In particular we discuss the applicability of different classifiers in a sparse representation domain and give insights into the classifier overtraining problem and biased error estimates with respect to varying training set sizes and the high dimensionality of data.

4.1 Introduction

As described earlier, generally the face recognition task is a two-stage process: a face detection/localization stage consisting of a fast detector and a classification/recognition stage using a complex classifier. A face recognition system should be able to deal with significant changes in the appearance of a single face. Within-class variations due to pose and viewing direction are almost always larger than between-class variations due to change in face identity. Two issues are central; the first is what features to use to represent a face. An effective representation should ideally be invariant to in-class variations, but distinctive with respect to face identity. The second issue is how to classify a new face image using the chosen representation.

In order to perform multi-view face recognition (recognizing faces under different poses) it is generally assumed to have examples of each person in different poses available for training. The problem is solved form a typical machine learning point of view where each person defines one class. A classifier is then trained that seek to separate each class by a decision boundary. Multi-view face recognition can be seen as a direct extension of frontal face recognition in which the algorithms require gallery images of every subject at every pose [B96]. In this context, to handle the problem of one training example, recent research direction has been to use specialized synthesis techniques to generate a given face at all other views and then perform conventional multi-view recognition [LK06][GMB04].

4.1.1 Performance considerations in multi-view face recognition

Here we focus on studying the effects on classification performance when a proper alignment is not assumed and there exist large pose differences. With these goals in mind, the generalization ability of classifier is evaluated with respect to the small sample size problem. Small sample size problem stems from the fact that face recognition typically involves thousands of persons in the database to be recognized. Since multi-view recognition treats each person as a separate class and tends to solve the problem as a multi-class problem, it typically has thousands of classes. From a machine learning point of view any classifier trying to learn thousands of classes requires a good amount of training data available for each class in order to generalize well. Practically, as discussed in previous chapters, we have only a small number of examples per subject available for training and therefore more and more emphasis is given on choosing a classifier that has good generalization ability in such sparse domain.

The other major issue that affects the classification is the representation of the data. The most commonly used feature representations in face recognition have been introduced in chapter 3. Among these the Eigenface by using PCA is the most common to be used in multi-view face recognition. The reason for that is the associated high dimensionality of other feature descriptions such as Gabor, LBPH etc. that prohibits the

use of any learning to be done. This is the well known curse of dimensionality issue in pattern recognition [DHS01] literature and this is just the reason that methods using such over complete representations normally resort to performing a simple similarity search by computing distances of a probe image to each of the gallery image in a typical template matching manner. While by using PCA on image pixels an upper bound on the dimensionality can be achieved.

In line with the above discussion, we therefore demonstrate the effectiveness of the proposed face-GLOH signatures with that of using conventional PCA based features in multi-view face recognition scenarios. Here we choose two of the well known face databases, the ORL face database and the FERET database, in order to assess the performance with regards to misalignments and pose variations. ORL database is chosen as many of the previous methods report near perfect performance by using conventional classifiers. In [SH97], a hidden Markov model (HMM) based approach is used, and the best model resulted in an error rate of 13%. Later, they extend the top-down HMM with pseudo two-dimensional HMMs reducing the error rate to 5%, [LG97] takes the convolutional neural network (CNN) approach for the classification of ORL database, and the best error rate reported is 3.85%. Whereas [GLK00] reports an average minimum error rate of 3.0% by using a SVM trained on 50% of the dataset and remaining for testing.

We show, in an extensive experimental setting, that even better classification scores can be obtained by carefully choosing the feature representation and using rather simpler classifiers. One of the contributions is the finding that these results are rather insignificant in the sense that problem is too simple: if simple even linear classifiers suffice, there is no sense in evaluating more complex and sophisticated classifiers on ORL database. For a more practical real world problem, because of the fact that detected faces are artificially aligned, we show that by introducing an arbitrary scale and shift variation in each of the example images the recognition results with almost all of the classifiers are severely affected for the conventional feature representation of the face.

ORL face database is mainly used to study the effects on classification performance due to misalignments since variations due to pose are rather restricted (not more than

20^{o}). To study the effects of large pose variations and a large number of subjects, we therefore repeat our experiments on FERET database.

In the next section we introduce the commonly used classifiers in the context of face recognition.

4.2 Linear and non-Linear Classifiers

Generally classifiers can be partitioned into linear classifiers and non-linear classifiers. Linear classifiers implement a linear decision boundary to discriminate classes. Some of the most widely used and bench-marked linear classifiers in face recognition includes the linear discriminant classifier (LDC), fisher discriminant analysis, nearest mean classifier (NMC) and support vector machine (SVM) classifier. While classifiers like quadratic discriminant classifier (QDC), non-parametric k-nearest neighbour (k-NN), the decision tree classifier and kernel based Parzen density estimator are examples of non-linear classifiers.

In addition to these, several classifier combining schemes have been proposed in the literature [SK02] [D02] [K02]. Several classifiers can be combined using fixed rules or using a trainable combiner on single or several representation sets. A complete and in-depth discussion of all of the classifiers mentioned above can be found in [W02]. Here we give a brief overview of these classifiers, since we are using these in our experiments.

4.2.1 Normal density-based classifiers

The most widely used classifiers are based on normal densities. Classification is achieved by assigning a pattern x to class w_i for which the posterior probability $p(w_i|x)$ is largest by assuming normal density for class conditional densities $p(x|w_i)$. The discriminant rule is to assign x to w_i if $g_i > g_j$ for all $j \neq i$, where

$$g_i(x) = \log(p(w_i)) - \frac{1}{2}\log(|\Sigma_i|) - \frac{1}{2}(\chi - \mu_i)^T \Sigma_i^{-1}(\chi - \mu_i) \qquad (4.1)$$

The parameters mean μ and covariance $\sum$ are the maximum likelihood estimates for each class, which when inserted in Equation (4.1) gives the Gaussian or quadratic QDC classifier.

LDC assumes the class covariance matrices $\sum_i,\ldots,\sum_c$ for C classes are equal, which simplifies Equation (4.1) and makes discriminant linear

$$g_i(x) = \log(p(w_i)) - \frac{1}{2}\mu_i^T S_W^{-1}\mu_i + x^T S_W^{-1}\mu_i \tag{4.2}$$

where S_W is the within class group covariance matrix also termed as within class scatter. For a total of 'N' samples in training data this scatter can be found simply by

$$S_W = \sum_{i=1}^{C} \frac{N_i}{N}\sum_i \tag{4.3}$$

When the Scatter matrix S_W is taken to be identity and the class priors $p(w_i)$ are equal, discriminant in Equation (4.2) reduces to the NMC, each test point is assigned the class label of the class whose mean is nearest.

4.2.2 Non-parametric and kernel-based classifiers

The k-NN, Parzen density estimator and SVM all come under this category. k-NN is a simple method of density estimation in which the probability of a test point x' falling within a volume V centered at some point x is approximated by the proportion of samples K falling within V

$$\mathrm{p}(\mathrm{x}) = \frac{K}{\mathrm{nV}} \tag{4.4}$$

The k-NN approach is to fix k and determine the volume V that contains k samples centred on point x. The decision rule is assign x to the class that yields the smallest volume for the k nearest neighbours.

The kernel variant of this method, namely the parzen density estimator, fixes the volume and finds the number of samples within the volume to estimate density.

$$p(x) = \frac{1}{hn}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{4.5}$$

Where *K(:)* is the kernel function, the most widely used is the normal form, which we also used in our experiments. h, (the proportion of observations falling within a fixed interval) is a smoothing parameter which determines the shape of the density. An optimal value can be determined during training to give an optimal width for each class separately.

4.2.2.1 Support vector machine

SVMs have demonstrated excellent performance for recognition tasks in computer vision [SS02]. They perform classification between two classes by finding a decision surface that maximizes the distance to the closest points in the training (the so called support vectors). For a binary classification problem a set of training vectors belonging to two separate classes $(x_1,y_1),\ldots..,(x_i,y_i)$, where $x_i \in R_n$ and $y_i \in \{1,-1\}$, are separated by a hyper-plane($w_x + b=0$) with maximum margin called optimal separating hyper-plane (OSH). The OSH is obtained by minimizing an objective function of the form

$$\phi(w,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{k}\xi_i \tag{4.6}$$

subject to constraints $y_i\left[\left(w \cdot x_i\right)+b\right] \geq 1-\xi_i$, $\xi_i \geq 0$ for $i=1,...,k$, where C is the regularization constant and ξ_i are the slack variables introduced to penalize errors when data are not linearly separable. The whole formulation is a constrained optimization problem. A test point x can now be classified by using the sign of the OSH decision surface function.

$$f(x) = \sum_{i=1}^{k} y_i \alpha_i K(x, x_i) + b \tag{4.7}$$

where $\alpha_i \geq 0$ are the Langrangian multipliers corresponding to support vectors, and b is found by solving the above mentioned optimization problem. K(.,.) is the kernel trick used to transform data onto a higher dimensional space where initially non-separable data can then be separated linearly by a hyper-plane. The most popular choices for the kernel function are the polynomial and Gaussian radial basis function (RBF) kernel. The

polynomial kernel is of the form $k(x,y)=(1+x\cdot y)^d$, d specifies all the possible monomials of the input components up to degree d. The RBF kernel is of the form $K(x,y)=\exp\left(-|x-y|^2/\sigma^2\right)$.

The whole formulation can be extended for the multiclass SVM classification by using one of two strategies. In the one-vs-all approach, m classifiers are trained for an m-class problem. Each of the SVM separates single class from the rest. In the pair-wise approach m(m-1)/2 SVMs are needed. Each SVM is trained on a pair of class and then all SVMs are arranged in a tree structure. There is no theoretical analysis of the two strategies with respect to classification performance. Regarding the training effort, the one-vs-all approach is preferable since only m SVMs have to be trained compared to m(m-1)/2 in the pair-wise case. Since the number of classes in face recognition can be rather large we opted for the one-vs-all strategy where the number of SVMs is linear with the number of classes.

4.2.3 Classifier combining

Classifier combining is a well studied subject and is used to improve classifier performance. Usually parallel and stacked combining are used. Stacked combining typically applies to different classifiers for the same feature space, while parallel classifier combiner uses different feature spaces. Combination is achieved using either fixed rules [D02] or the combiner itself can be trained using the training set. The details involved are an ongoing subject of the research [K02] [SK02]. An in-depth discussion on combining classifiers can be found in [K04].

In our experiments, we chose to test several fixed rules for combining using both parallel and stacked combining schemes. We use 6 fixed combining rules, namely median, maximum, minimum selection as well as mean, product and voting combiner.

As noted in [D02], it is necessary to scale the outputs of different classifiers in a way such that the outputs become comparable. For our experiments the output of different classifiers is normalized by fitting a sigmoid on the outputs such that the sum of the outputs becomes 1. This out put normalization is necessary for two reasons. Firstly

for fixed combining rules, comparable outputs of base classifiers are needed and secondly output normalization may inject non-linearity and hence significantly improve the combiner performance.

4.3 Experimental Setup

As discussed before, we perform experiments in order to assess the performance of various classifiers in a typical multi-view recognition scenario with respect to the following factors.

- Feature representations
- When facial images are not artificially aligned
- When there are large pose differences
- Large number of subjects
- Number of examples available in each class (subject) for training.

In order to show the effectiveness of face-GLOH signature feature representation against misalignments, we use ORL face database. ORL face database has 40 subjects depicting moderate variations among images of same person due to expression and some limited pose. However, all the images are depicted in approximately the same scale and thus have a strong correspondence among facial regions across images of the same subject. We therefore generate a scaled and shifted ORL dataset by introducing an arbitrary scale change between 0.7 and 1.2 of the original scale as well as an arbitrary shift of 3 pixels in random direction in each example image of each subject. This has been done to ensure having no artificial alignment between corresponding facial parts. This new misaligned dataset is denoted scaled-shifted SS-ORL (see Figure 4-1). The experiments are performed on both the original ORL denoted O-ORL and SS-ORL using PCA based features and face-GLOH signatures.

Next we use FERET database pose subset (comprising 200 subjects in 9 pose variations), as introduced in chapter 2. Experiments on FERET are performed in order to assess the performance with regards to increasing number of subjects and large pose variations. All the images are cropped from the database by using standard normalization methods i.e. by manually locating eyes position and warping the image onto a plane

where these points are in a fixed location. The FERET images are therefore aligned with respect to these points. This is done in order to only study the effects on classifier performance due to large pose deviations. All the images are then resized to 92x112 pixels in order to have the same size as that of ORL faces. An example of the processed images of a FERET subject depicting all the 9 pose variations is shown in Figure 4-2.

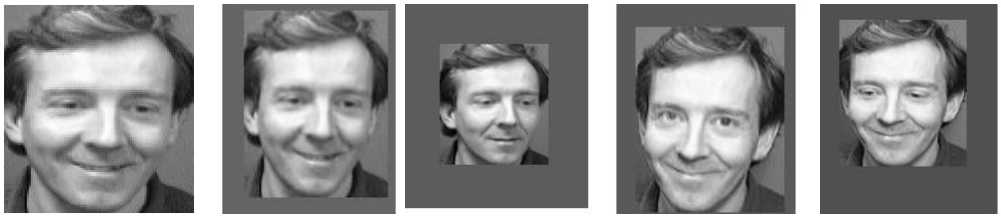


Figure 4-1: An example of a subject from O-ORL and its scale and shifted examples from SS-ORL

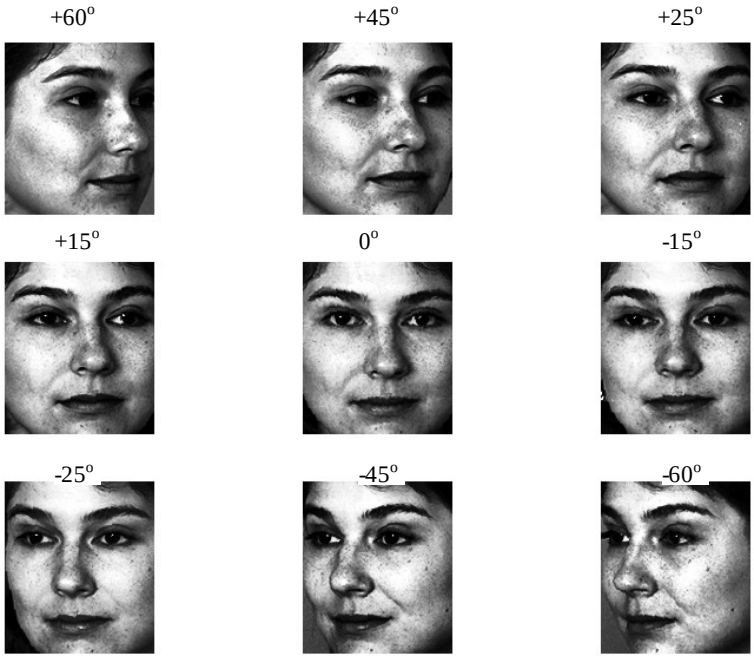


Figure 4-2: Cropped faces of a FERET subject depicting all the 9 pose variations.

4.3.1 Feature Extraction

For our first representation we extract one global feature vector per face image by using lexicographic ordering of all the pixel grey values. Thus, for each 92 x 112 image, one obtains a 10384 dimensional feature vector per face. We then reduce this dimensionality by using unsupervised PCA. Where the covariance matrix is trained using 450 images of 50 subjects from FERET set. The number of projection Eigen-vectors are found by analysing the relative cumulative ordered eigenvalues (sum of normalized variance) of the covariance matrix. We choose first 50 largest Eigen vectors that explain around 80% of the variance as shown in figure 4-3. By projecting the images on these, we therefore obtain a 50-dimentional feature vector for each image. We call this representation the PCA-set.

The second representation of all the images is found by using face-GLOH-signature extraction, as detailed in the chapter 3. An obtained face-GLOH-signature for a subject from ORL database is shown in figure 4-4.

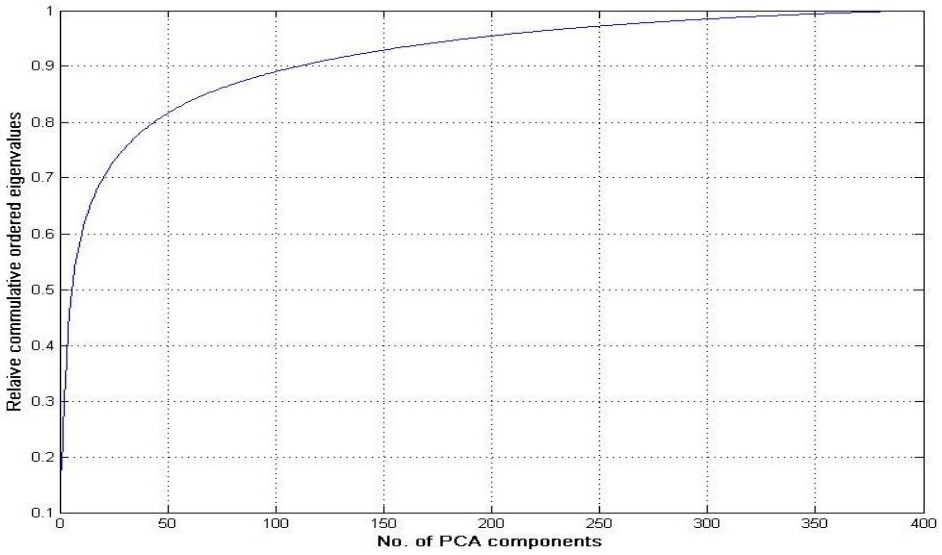


Figure 4-3: Plot of relative cumulative ordered eigenvalues for choosing PCA components.

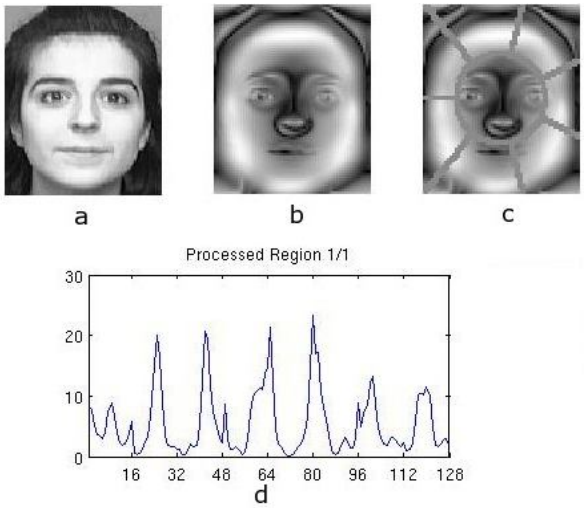


Figure 4-4: face-GLOH-signature extraction for a subject in O-ORL database

4.3.2 Classifier performance measure

The ultimate measure of a classifier performance is the classification error or simply the error rate (P_e). Competing classifiers can also be evaluated based on their error probabilities. Other performance measures include the cost of measuring features and the computational requirements of the decision rule [JDM00].

The performance measure P_e should be an unbiased estimate. The error estimate of a classifier, being a function of the specific training and test sets used, is a random variable. Given a classifier, suppose η is the number of test examples (out of a total of N) that are misclassified. It can be shown that the probability density function of η has a binomial distribution. The maximum-likelihood estimate, $\hat{P}_e$ of P_e is given by $\hat{P}_e = \eta / N$ with $E(\hat{P}_e) = P_e$ and $\text{var}(\hat{P}_e) = P_e(1-P_e)/N$. P_e, is therefore an unbiased and consistent estimator.

The classifier is first designed using training examples, and then it is evaluated based on its classification performance on the test examples. The percentage of misclassified test examples is taken as an estimate of the error rate. In order for this error estimate to be reliable, the training examples and the test examples must be independent.

This requirement of independent training and test examples is still often overlooked in practice. There are no good guidelines available on how to divide the available samples into training and test sets, if the training set is small, then the resulting classifier will not be very robust and will have low generalization ability. On the other hand, if the test set is small, then the confidence in the estimated error rate will be low. No matter how the data is split into training and test sets, it should be clear that different random splits (with the specified size of training and test sets) will result in different error estimates.

Keeping these factors in mind we have used two procedures to ensure a fair estimate of errors of all classifiers. The first is to use different varying training set sizes by using different number of examples per class and testing on the remaining. Tests at each size are repeated 5 times, with different training/test partitioning of the examples per subject/class, and the errors are averaged. The second uses a 10-fold cross validation procedure to produces 10 sets of the same size as original dataset each with a different 10% of objects being used for testing. All classifiers are evaluated on each set and the classification errors are averaged.

4.3.3 Experiments on ORL datasets

In all of our experiments we assume equal priors for training, SVM experiments on O-ORL use a polynomial kernel of degree 2, to reduce the computational effort, since using RBF kernel with optimized parameters C and kernel width σ did not improve performance. For SS-ORL a RBF kernel is used with parameter C=500 and $\sigma = 10$, these values were determined using 5-fold cross validation and varying sigma between 0.1 and 50 and C between 1 and 1000. All the experiments are carried out for classifiers on each of two representations for both O-ORL and SS-ORL.

First set of experiment use 10-fold cross validation as described before. The results from this experiment on both O- ORL and SS-ORL for both feature representations are reported in table 4-1.

Table 4-1: Classification errors in 10-fold cross validation tests on ORL

Classifiers	O-ORL Representation sets		SS-ORL Representation sets	
	PCA	face-GLO H	PCA	face-GLOH
NMC	0.137	0.152	0.375	0.305
LDC	**0.065**	**0.020**	0.257	0.125
Fisher	0.267	0.045	0.587	**0.115**
Parzen	0.037	0.030	0.292	0.162
3-NN	0.097	0.062	0.357	0.255
Decision Tree	0.577	0.787	0.915	0.822
QDC	0.64	0.925	0.760	0.986
SVM	**0.047**	**0.037**	0.242	**0.105**

Table 4-1 shows how classification performance degrades, when the faces are not aligned i.e. arbitrarily scaled and shifted, on PCA based feature representation. The robustness of the face-GLOH-signature representation against misalignments can be seen by comparing the results on O-ORL and SS-ORL, where it still gives comparable performance in terms of classification accuracy. Best results are achieved by using LDC or SVM in both cases.

The second set of experiments uses varying training set sizes of 2,4,6 and 8 examples per subject and testing on the remaining, tests at each size are repeated 5 times, with different training/test partitioning, and the errors are averaged. The resulting error curves for SS-ORL and O-ORL are shown in figure 4-5. This experiment specifically address the problem of biased error estimates with respect to varying training and test sizes of the dataset. Figure 4-5 (b) and 4-5 (d) shows the effect on classifier performance with respect to misalignments and the advantage of a better feature representation.

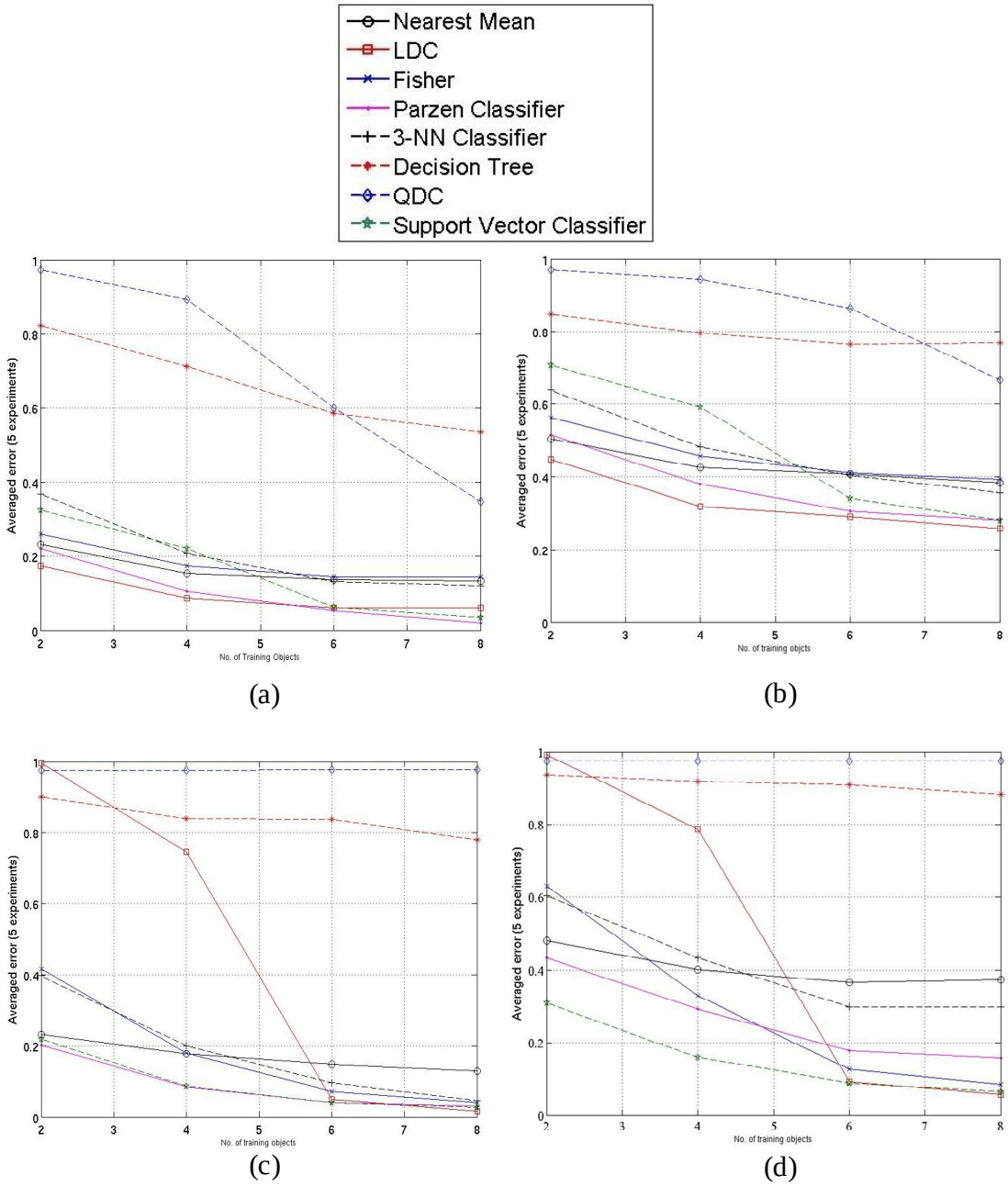


Figure 4-5: Classifiers evaluation by varying training set sizes (a) On O-ORL using PCA-set (b) On SS-ORL using PCA-set (c) On O-ORL using face-GLOH-signature set (d) On SS-ORL using face-GLOH-signature set.

4.3.4 Experiments on FERET dataset

As described before, 50 out of 200 FERET subjects are used for training the covariance matrix for PCA. The remaining 1350 images of 150 subjects are used to evaluate classifier performance with respect to large pose differences. Following the same procedure, experiments on FERET set are performed with respect to varying training/test sizes by using 2, 4, 6, and 8 examples per subject and testing on the remaining. Similarly, tests at each size are repeated 5 times, with different training/test partitioning, and the errors are averaged. Figure 4-6 shows the averaged classification errors for all the classifiers on FERET set for both the feature representations with respect to varying training and test sizes. As shown in figure 4-6, increasing number of subjects and pose differences has an adverse affect on the performance of all the classifiers as compared to ORL.

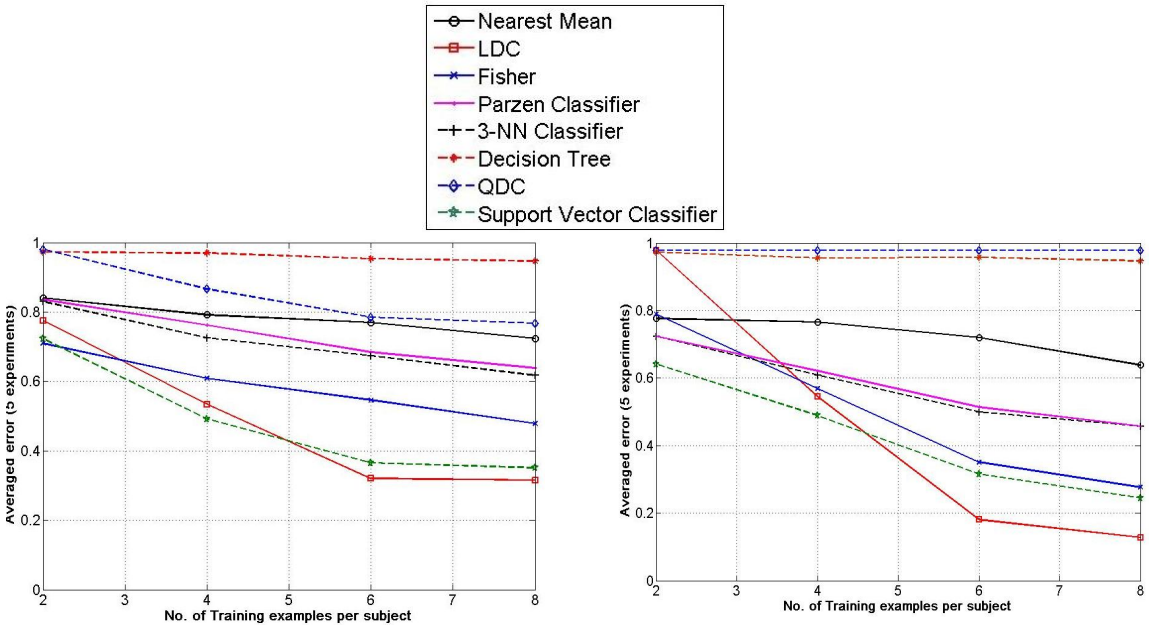


Figure 4-6: Classifiers evaluation On FERET by varying training/test sizes

(a) Using PCA-set (b) Using face-GLOH-signature set

Table 4-2: Classification errors in 10-fold cross validation tests on FERET

Classifiers	FERET Representation sets	
	PCA	face-GLOH-signature
NMC	0.789	0.703
LDC	**0.465**	0.399
Fisher	0.576	0.475
Parzen	0.735	0.553
3-NN	0.815	0.673
Decision Tree	0.881	0.852
QDC	0. 751	0.916
SVM	0.550	**0.297**

Table 4-2 lists the classification errors in a 10-fold cross validation setting on FERET by using each of the two feature representations. As shown, the best score on PCA based representation is achieved by LDC i.e. 46.5% that amounts to a classification accuracy of 53.5%, where as a significant improvement is observed for the classification scores of almost all the classifiers on face-GLOH-signature representation. The best classification accuracy is about 70.3% by using SVM on face-GLOH-signature representation. These results indicate the advantage of using a more robust feature representation as opposed to commonly used PCA based features in a typical multi-view face recognition task.

4.3.4.1 Results using classifier combining

As compared to the results on ORL database, the effect, of large pose differences and large number of subjects, on the classification performance is quite severe. A drop of about 20 % to 30% of performance is observed on FERET for the best performing classifiers including LDC and SVM as compared to the results on O-ORL and SS-ORL datasets. In [SJH06] we have proposed to use classifier combining to further improve the results. We first explore the use of six fixed combining rules on each of the feature set

(stacked combining) by using two simple LDC and fisher as base classifiers. The results (table 4-3) indicate the significant performance gain on both of the feature sets. Further, we use LDC as base combiner on both of the feature set. This experiment specifically explores the benefit of using multiple feature representation for the same object, in our case the faces. We have used 50% (randomly chosen) of the data for training the classifiers and rest for testing. This amounts to having 5 examples per subject for training and 4 for testing. Compared to the results reported in figure 4-6 for 5 training examples/objects per subject, the classifier combining improves the classification accuracy significantly.

Table 4-3: Classification scores by combining classifiers on FERET

Classifier combiners	**Stacked over each of the feature set using LDC and fisher as base classifier**		**Parallel on both feature sets**
	PCA	Face-GLOH-signature	Using LDC as base combiner on both representation sets
Product combiner	**0.223**	0.266	0.207
Mean combiner	0.531	**0.213**	**0.196**
Median combiner	0.531	0.213	0.196
Maximum selection	0.491	0.218	0.197
Minimum selection	0.286	0.396	0.209
Voting combiner	0.383	0.388	0.218

As can be seen from the results in table 4-3, best scores in stacked combining are achieved for product and mean combiner. The classification errors dropped from 44.5 % to 22.3% on PCA set, and from 39.9% to 21.3% on face-GLOH-signature set. These results by combining two simple classifiers are significantly better than by suing a more complex classifier such as SVM.

Parallel combining on both the feature sets using LDC as base combiner achieves the best result i.e. 19.6% classification error, a drop of about 20% classification error when compared to the best performing classifier on each of the representation set by using same number of training and test examples per subject (see figure 4-6). These results are much better than the best performing SVM (29.7% classification error) as reported in table 4-2. Overall classification accuracy improves form about 68% to 81.4% by using classifier combing on FERET.

4.4 Conclusion

In a typical multi-view face recognition task, where it is assumed to have several examples of a subject available for training, we have shown in an extensive experimental setting the advantages and weaknesses of commonly used classification methods. Our results show that under more realistic assumptions, most of the classifiers failed on conventional features. While using the introduced face-GLOH-signature representation is relatively less affected by large in-class variations. This has been demonstrated by providing a fair performance comparison of several classifiers under more practical conditions such as misalignments, large number of subjects and large pose variations.

The reason for a good performance for simple linear classifiers on O-ORL database lies in the fact that the problem is very sparse and the classes are linearly separated in such high dimensions, but as shown by our results on scaled and shifted dataset, a more robust feature representation is needed to cater the large in-class variance. One reason for poor performance of QDC as compared to others is that in such high dimensions with too few objects, there is less hope to estimate any density based on parametric methods. For such problems a regularization imposed on the covariance matrix is needed.

Our results on FERET indicate a consistent performance degradation of all the classifiers. As compared with ORL our experiments confirm the fact that to achieve good generalization performance of classifiers, one needs a large and more difficult dataset. Classifier combining does improve the classification, for our problem fixed combining on all feature sets has produced appreciable improvements in results.

An important conclusion is to be drawn from the results on FERET is that conventional multi-view face recognition can not cope well with regards to large pose variations. Even using a large number of training examples in different poses for a subject do not suffice for a satisfactory recognition. In order to solve the problem where only one training example per subject is available, many recent methods propose to use image synthesis to generate a given subject at all other views and then perform a conventional multi-view recognition[BP95][GMB04]. Besides the fact that such synthesis techniques cause severe artefacts and thus can not preserve the identity of an individual, a conventional classification can not yield good recognition results, as has been shown in an extensive experimental setting. More sophisticated methods are therefore needed in order to address pose invariant face recognition. Large pose differences cause significant appearance variations that in general are larger than the appearance variation due to identity. One possible way of addressing this is to learn these variations across each pose, more specifically by fixing the pose and establishing a correspondence on how a person's appearance changes under this pose one could reduce the in-class appearance variation significantly, we will elaborate more on that in later chapters. The pose information is valuable and for a fully automatic face recognition system the pose of the incoming probe/test image has to be known in this context. An effective pose estimation framework for the purpose of automatic face recognition is the subject of next part of this thesis.

Next we plan to use some of these insights in developing a more generic face recognition framework that is able to recognize faces from just one training image per subject. We hope that a fair comparison, with respect to carefully chosen training and test sizes and different feature representations, gives a useful review and guidance while developing different face recognition tasks.

PART - II

On Head Pose Estimation in Face Recognition

Chapter 5. Head Pose Estimation for Automatic Face Recognition

In this chapter we introduce the problem of head pose estimation in the context of face recognition. Facial pose plays an important role while recognizing faces in different poses. Besides face recognition, head pose estimation has other many useful applications in different automated vision tasks. Here we provide a brief discussion on associated problems in pose estimation and the existing pose estimation methods in the context of face recognition. Furthermore, this chapter introduces a novel feature description termed as LESH (Local Energy based Shape Histogram) that encodes the underlying shape in an illumination and skin texture invariant manner for the purpose of face pose estimation.

5.1 Background

The primary concern of this thesis is recognizing persons in the presence of large pose and illumination variations. When considering pose variations, the test input image can be transformed to all the reference poses before performing face recognition [BP95] or else more preferably the problem can be approached by first estimating the pose of the test input image and then transform it to a reference pose where face recognition can be performed at that pose or the top few poses [LK06] [CSB06]. In the context of automatic face recognition, we therefore need a front-end face pose estimation system that is able to estimate the pose of an incoming test input face image in some discrete pose labels.
Besides this head pose estimation finds many useful applications in other computer aided vision tasks e.g. human computer interaction [FH07] [MSLD07], driver assistance [CPT06] [CT08] etc.

In the context of computer vision, head pose estimation is most commonly interpreted as the ability to infer the orientation of a person's head relative to the view of

a camera. At the coarse level, head pose estimation applies to algorithms that identify a head in one of a few discrete orientations, e.g., a frontal versus left/right profile view. At the fine (i.e., granular), a head pose estimate might be a continuous angular measurement along the rotation of head in multiple Degrees of Freedom (DOF) as is done in head tracking systems [ZCSC07]. The human head is limited to 3 DOF in pose, which can be characterized by pitch, roll, and yaw angles as illustrated in Fig. 5-1.

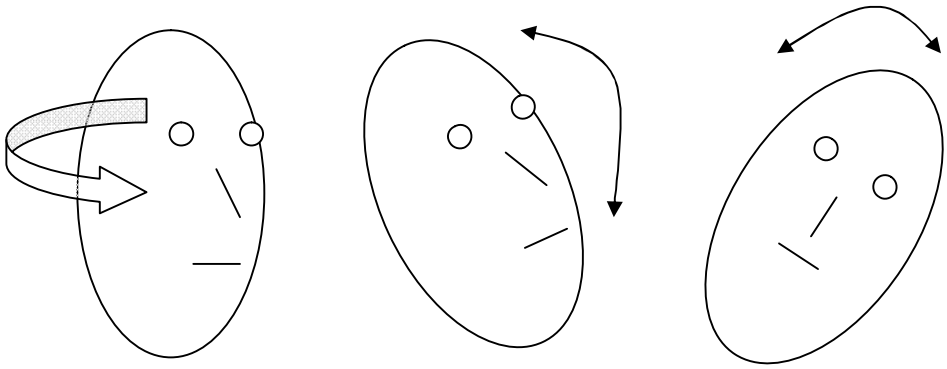


Figure 5-1: Head Orientation in 3DOF

Yaw describes the in-depth rotation of the head relative to the camera. Pitch corresponds to up and down tilt of the face and roll is the side tilt of the face. It is common to use the geometric normalization based on eyes location to remove this tilt bias of the face by rotating the face such that both the eyes lies on the same horizontal line. The variations in facial appearance caused by yaw are much greater than variations by either pitch or roll. Some face representation or feature extraction techniques may provide inherent invariance to some degree towards pitch or roll of the face e.g. Gabor, LBPH or our introduced face-GLOH-signature. However, it is not directly possible to extract invariant features with respect to yaw rotation, this is because of the fact that yaw or in-depth rotation of the face results in a significant loss of information, e.g. when a face turns

from frontal to full left or right profile view almost half of the face is not visible and the position and appearance of the facial parts looks vastly different. Pose estimation, therefore, refers to estimating this in-depth rotation (yaw) of the face in order to use this information for the benefit of face recognition.

5.1.1 Related work

Head pose estimation has become a broad area of research over the recent years. Generally, methods for face pose estimation can be categorized into model-based approach or appearance based methods [BT02].

Model-based algorithms are based on feature detection and usually need several manually localized landmark points on the face and use this information to model the displacement of these points with respect to different poses. Among these are Active Appearance Models [MB04] that are nonlinear parametric models derived from linear transformations of a shape model and an appearance model, as discussed in chapter 2. Also, Neural Network based algorithms can be trained to distinguish between different persons or to distinguish poses of one persons face [RBK98].

Gee and Cipolla [GC94] chose a few relatively stable feature points, or anchor points, to estimate the gaze direction under weak perspective, with an assumption that the ratios of these points did not change significantly for different facial expressions. Gao et al. [GLWH01] presented an efficient pose recovery approach using locations of the two eyes and the symmetric axis of the face, but their method is very sensitive to the location error of the facial features. A statistical 3D morphable model [BV03] is proposed to deal with pose variations as well as illuminations, which requires textured 3D scans of heads and involves expensive computations. Dias and Buxton [DB04] proposed an integrated model by combining Active Shape Models [CT95] with Ullman and Basri's Linear Combination of Views (LCV) [UB97]. The LCV technique, however, requires accurately locating at least three corresponding feature points in each of the 2D images of different views.

Appearance based approaches avoids the expensive feature localization and tend to learn the whole appearance of a face in different poses. PCA is typically used by finding a subspace where pose variation can be described linearly. Many subspace-based

techniques have been used to tackle pose-invariant face recognition. Pentland et al used a view-based subspace method by producing separate subspaces each constructed from faces at the same viewpoint [PMS94]. Murase and Nayar used a parametric eigen-space method by representing each known person by compact manifolds in the subspace of the eigenspace [MN95], recognition is performed by first finding the subspace most representative of the test face and then matching using a simple distance metric in this subspace. Analytical subspace method is used by Valentin and Abdi [VA96]. Characteristic subspace method is used by McKenna et al [MGC96]. PCA based algorithms in general assume that appearance variation due to pose are larger than appearance variations due to subject identity. This assumption in general is not true since different subjects across same pose may have large appearance variations due to e.g. glasses, expressions, illumination and skin color. To overcome this some recent appearance-based approaches use filtering and image segmentation techniques to extract information from the image. In [SGO99] [WFT01] authors use Gabor response to construct the eigenspaces, [KS02] used Gabor wavelet network to estimate pose. Similarly [PC05] construct a filter trained on different poses for later recognition of facial pose.

Recently some hybrid methods have been presented that combine the appearance based approaches and model based methods. Gründig and Hellwich [GH04] use 3 landmarks detected on the face for initial pose estimate and combine it with PCA based Eigen basis functions to refine the pose estimate. In [CT08] authors combine PCA embedded template matching with optical flow, [CT04] used PCA based template matching with a continuous density hidden Markov model. A more comprehensive account of various head pose estimation approaches can be found in a very recent survey [CT08b].

5.1.2 Concerns in pose estimation

Most of the existing approaches for head pose estimation make some strong assumptions with regards to the problem at hand that are very hard to meet in practice. Model based approaches, for example, requires several detected landmark points in order to estimate pose, they therefore are very sensitive with the localization of these points and as

discussed in chapter 2, an automatic localization of these points is not trivial. Besides this, the general assumption of modeling the displacement of these points across different poses is based on a very strong assumption of having much less shape variation within same pose (intra-pose) than among different pose (inter-pose). Appearance based methods on the other hand, although avoids these problems of landmarks localization and modeling, but it also assumes that inter-pose appearance variations are always larger than intra-pose appearance variations. This, generally, is not true since different subjects across same pose may have large appearance variations due to e.g. glasses, expressions, illumination and skin color. It is therefore not easy to discriminate among the shape variations due to pose and variations due to identity (different faces). Illustration in Figure 5-2, adapted from [ZG06], serves to illustrate this point.

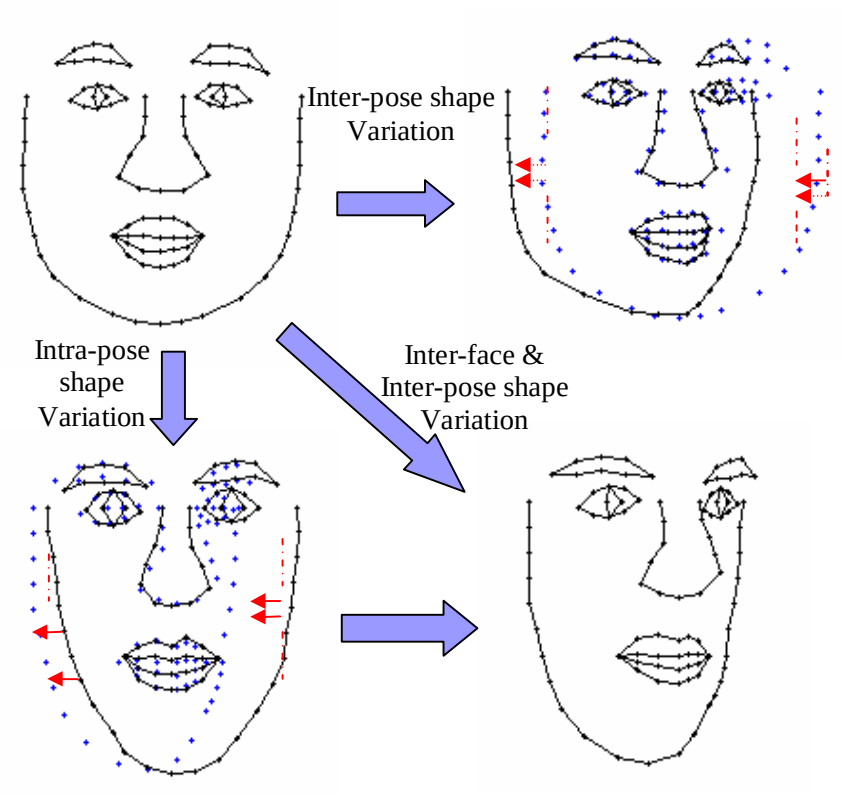


Figure 5-2: Inter-pose and Intra-pose shape variations

As shown in figure 5-2, these shape variations have to be taken into account while designing any pose estimation system. In order to have a fully automatic system appearance based methods are preferable since they avoid an expensive localization/detection of landmark points. To overcome problems with regards to appearance variation in same pose due to different subjects, ideally one needs an identity invariant representation that is invariant to skin texture and other subject specific variations. This is in contrast with the face recognition paradigm where the goal of common feature representations is to preserve the identity.

Keeping in view these factors, we propose an appearance-based head pose estimation method in this part of the thesis that overcomes some of these problems and addresses issues such as, a suitable representation and efficient estimation procedures in the context of automatic face recognition.

5.2 A New Feature Description for Pose Estimation

In this section we propose a new feature description that is based on a pure shape representation. This is in-line with our preceding discussion that any feature description or representation for the purpose of pose estimation should be invariant to subject identity i.e. skin texture, color etc. In current appearance based methods, as mentioned before, image filtering such as using a Gabor wavelet transform is used in order to enhance the image shape information such as edges etc. Recently, Baochang et al, [BSXW07] has shown that local Gabor phase patterns can provide a very informative description. This however, is still dependant on the person specific appearance variations, since multi resolution Gabor filtering results in a highly localized response at each image position dependant on the underlying surface properties. Despite this, such representations are sensitive to illumination variations present in images. In the context of unconstrained face recognition, the pose estimation has to be robust with regards to the lighting variations. Nonetheless, in a recent work (Bingpeng et al, [BWSXW06]) has shown that Gabor phase response is quite useful in order to model the orientation of the head.

Another body of work exists which uses this phase information to compute the local energy content of the underlying signal, for detecting interest points such as corners, edges, or contours etc.. We introduce a novel feature descriptor LESH (Local Energy based Shape Histogram) that is based on this local energy model of feature perception.

5.2.1 Local energy model

The local energy model developed by Morrone and Owens [MO87] postulate that features are perceived at points in an image where the local frequency components are maximally in phase. Morrone and Owens define the phase congruency function in terms of the Fourier series expansion of a signal at some location x as

$$E(x) = \max_{\bar{\phi}(x) \in [0, 2\pi]} \frac{\sum_n A_n \cos(\phi_n(x) - \bar{\phi}(x))}{\sum_n A_n} \tag{5.1}$$

Where A_n and ϕ_n are the magnitude and phase of the nth Fourier component. This frequency information must be obtained in a way such that underlying phase information is preserved. For this linear phase filters must be used in symmetric anti symmetric pair. This is achieved by convolving the image with a bank of Gabor wavelets kernels tuned to 5 spatial frequencies and 8 orientations. At each image location, for each scale and orientation, it produces a complex value comprising the output of even symmetric and odd symmetric filter, which gives the associated magnitude and phase of that pixel.

$$\mathrm{G}(\mathrm{e}_{n,v}, \mathrm{o}_{n,v}) = I(x,y) * \psi_{n,v}(z) \tag{5.2}$$

Where $\psi_{n,v}$ is the bank of Gabor kernel, n, v is the scale and orientation and $G(\cdot)$ is the response at image position (x, y) having a real and imaginary part comprising output of even symmetric and odd symmetric filter at scale n and orientation v. The amplitude A_n and phase ϕ_n thus can be written in terms of these responses at a given scale n.

$$A_n = \sqrt{e_n^2 + o_n^2} \qquad \textit{and} \qquad \phi_n = \tan^{-1}\frac{e_n}{o_n} \tag{5.3}$$

Originally [RO97] has proposed to use cosine of the deviation of each phase component from the mean phase as a measure of the symmetry of phase, however, this measure

results in poor localization and is sensitive to noise. Kovesi [K00] extended this framework and developed a modified measure, as given in equation 5.4, consisting of sine of the phase deviation, including a proper weighing of the frequency spread W and also a noise cancellation factor T.

$$E = \frac{\sum_n W(x)\left\lfloor A_n(x)(\cos(\varphi_n(x) - \bar{\varphi}(x)) - \left|\sin(\varphi_n(x) - \bar{\varphi}(x))\right|) - T \right\rfloor}{\sum_n A_n(x) + \varepsilon} \qquad (5.4)$$

The normalization by summation of all component amplitudes makes it independent of the overall magnitude of the signal, making it invariant to illumination variations in images. For details of this measure see [K00]. This illumination invariance property can be seen in Figure 5-3 that depicts the local energy response on an image that is very badly illuminated, the bright spots show the points of high local energy. As can be seen it is only high at edges, corners etc.

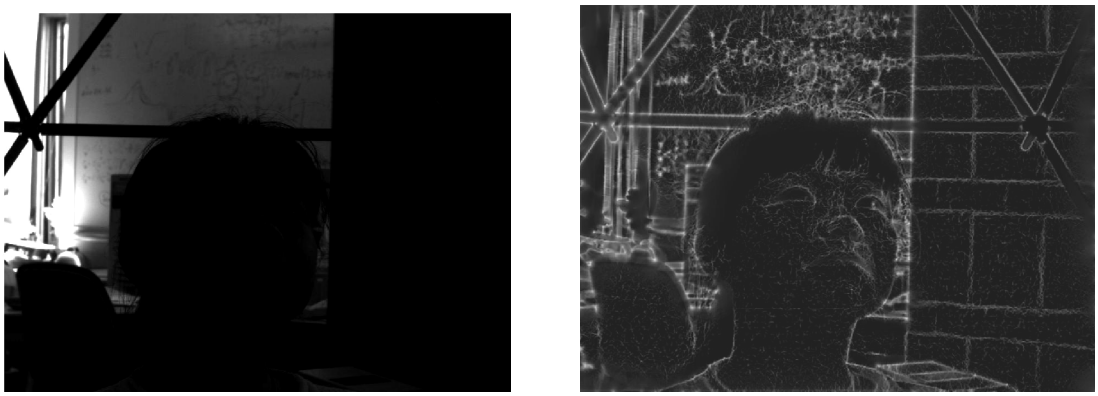


Figure 5-3: Local energy response for a badly illuminated image

5.2.2 LESH - Local Energy based Shape Histogram

The local energy analysis in the preceding section is intended to detect interest points in images with a high reliability in presence of illumination and noise. Hence, to detect these intrinsic two dimensional i2D structures. Kovesi [K03] proceeds by constructing principal moments of this normalized energy measure, also termed as phase congruency. In contrast to this, we rather use this raw energy information and attempt to encode the underlying shape. This is done in a way that makes it invariant to scale variations but not to rotation since rotation is precisely what, we are trying to model.

Motivated by the fact that this local orientation energy response varies with respect to the underlying shape, in our case the rotation of the head and since local energy signifies the underlying corners, edges or contours, we generate a local histogram accumulating the local energy along each filter orientation on different sub-regions of the image. The local histograms are extracted from different sub-regions of the image, and then concatenated together, to keep the spatial relationship between facial parts.

We proceed by obtaining an orientation label map where each pixel is assigned the label of the orientation at which it has largest energy across all scales. The local histogram 'h' is extracted according to the following

$$h_{r,b} = \sum w_r \times E \times \delta_{Lb} \tag{5.5}$$

where subscript 'b' represents the current bin, 'L' is the orientation label map, 'E' is the local energy, as computed in equation 5.4, δ_{Lb} is the Kronecker delta

$$\delta_{Lb} = \begin{Bmatrix} 1, & if & L = b \\ 0, & if & L \neq b \end{Bmatrix} \tag{5.6}$$

and 'w' is a Gaussian weighing function centered at region 'r'.

$$w_r = \frac{1}{\sqrt{2\pi\sigma}} e^{[(x - r_{xo})^2 + (y - r_{yo})^2]/\sigma^2} \tag{5.7}$$

This weight is used to provide soft margins across bins by small weighted overlap among neighbouring sub-regions to overcome the problems induced due to scale variations.

As mentioned earlier, in order to keep the spatial relation between facial parts, we extract 8 bins local histogram corresponding to 8 filter orientations on a 4x4 location grid (16 image partitions), which makes it a 128-dimentional feature vector. Figure 5-4 illustrates the schematic.

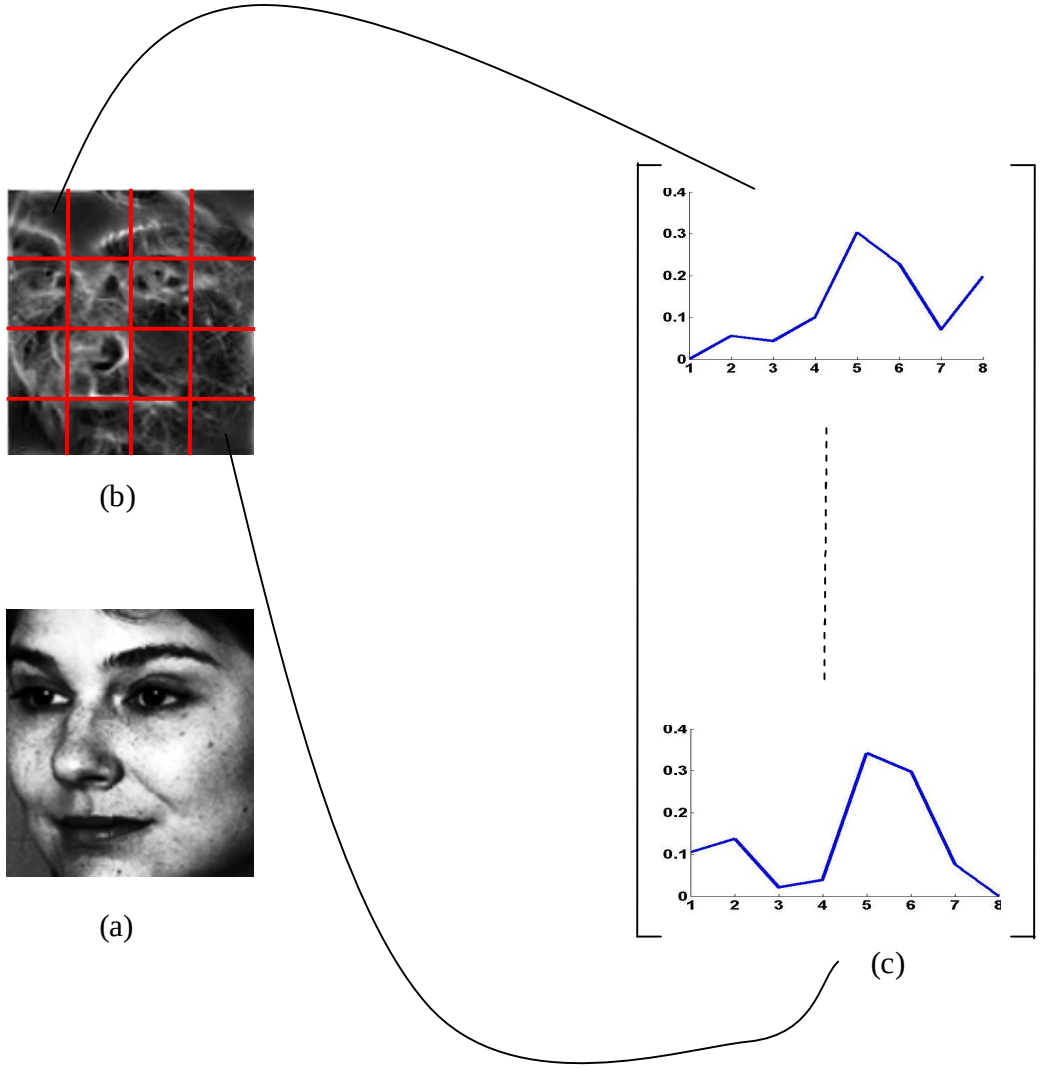


Figure 5-4: Schematic of LESH extraction: (a) Original image (b) 4x4 location grid imposed on corresponding Local energy map (c) 8-bin local histogram extracted from each partition and concatenated together into a 128-dimensional feature vector

Example feature extraction and associated energy and orientation maps on two different subjects in frontal and left profile pose, from CMU-PIE database, are shown in Figure 5-5. Figure 5-5 provides an intuitive look at the notion of similarity across same pose among different subjects, in terms of extracted local energy and LESH features. This notional similarity is validated empirically in chapter 6 by computing similarities between extracted LESH features. Note how they are quite invariant to person specific appearance variations.

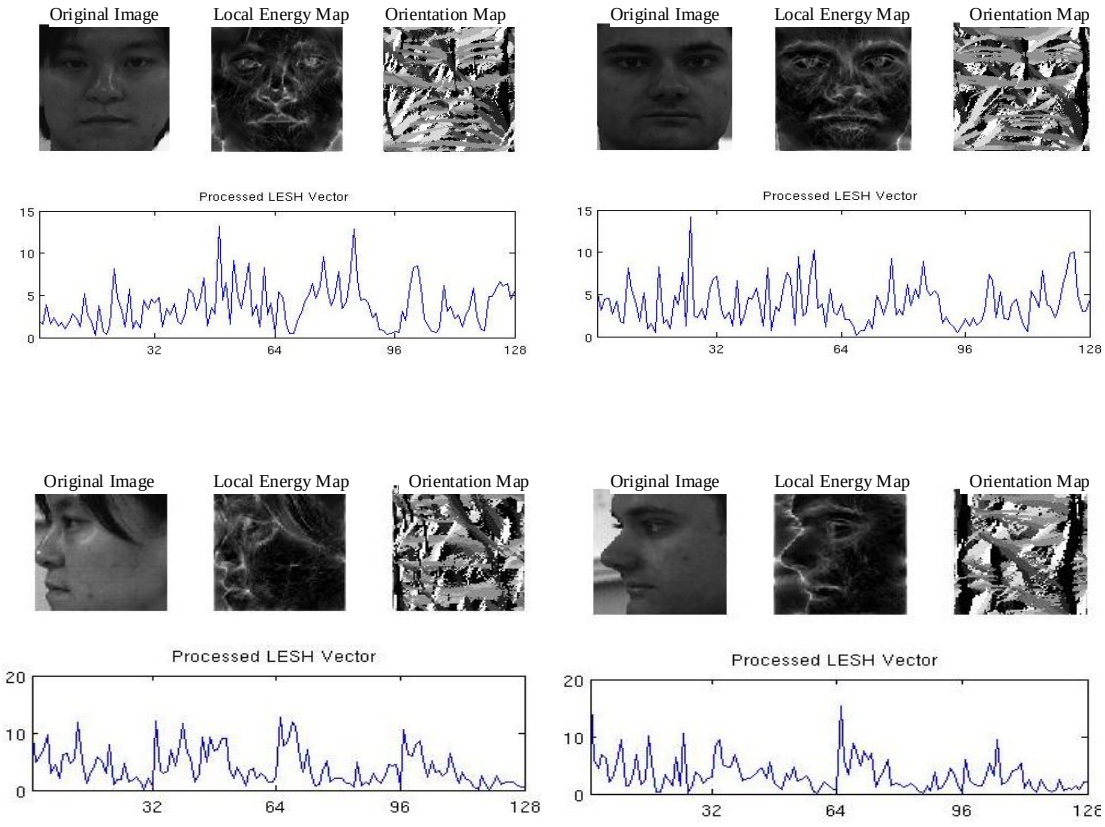


Figure 5-5: Two different subjects at frontal and left profile pose. Their associated energy and orientation maps and extracted LESH feature vectors.

5.3 Conclusion

In this chapter we have presented some background relating to head pose estimation in the context of face recognition. In addition to that, we have introduced a new feature description based on local energy model of feature perception, termed as LESH, which encodes the underlying shape and is insensitive to skin color and illumination variations. The effectiveness of LESH descriptor in the context of pose estimation will be shown in the next chapter.

Chapter 6. Head Pose Estimation Framework

The derived LESH features in the preceding chapter provide a strong foundation on which to base our pose estimation framework. These features are robust against slight misalignments and scale changes, but this comes with the cost of rather loose description of facial landmark positions. Although, this does not affect while discriminating among large pose variations such as between frontal and right or left profile view, but discriminating among nearest pose changes, by simply looking at similarity scores, may be quite error prone. Therefore, in order to discriminate between adjacent poses and to cater in-pose variations, due to other factors such as glasses, expressions and to some extent illumination (shadows), we need to learn these variations in a training phase.

We therefore, learn these variations across same pose from a training procedure in a novel way. In particular, we lay down an effective classification procedure that attempts to model these in pose variations and performs quite well in discriminating among slight pose variations.

6.1 Overview

In the context of face recognition pose estimation generally means identifying the pose of the incoming test image as one of some discrete pose labels. We therefore, can solve the pose estimation as a classification problem from a strict machine learning point of view. Pose estimation can be seen as a multi-class problem where the goal is to classify the test pose into one of the pose classes.

6.1.1 Estimation as a classification problem

Some recent works on estimating facial pose for later face recognition stage employ the same approach i.e. solve the pose estimation as a classification problem [PC05] [YC05] [BWSXW06]. All of these approaches tend to use a pure multi-class learning strategy where a classifier is trained on each of the predefined pose class. A test image is then assigned the label of the pose class for which the trained classifier gives the highest score. A fundamental problem with such a strategy is two-fold. On the one hand, fixing the discrete pose labels intrinsically requires any test image to be classified into one of these poses. In this context adding more pose classes to be recognized (depending on a given application), will require repeating the whole expensive training phase all over again. Secondly, usually such methods are discriminative in nature, which means, based on the classification result these systems return a hard pose label for any given test image.

From a practical stand point, the training should be an offline component, where it should be able to incorporate any new pose classes without the need of an expensive re-training phase. Moreover for a test image, instead of a black and white decision the classification method should provide the probabilities for each pose it might belong. This is more useful for the later face recognition stage, since this can be incorporated as a strong prior knowledge in the statistical sense. We will elaborate more on that in the next chapter.

Keeping in view these goals, in this chapter we propose an efficient learning procedure that turns this multi-class problem into a two-class one by modeling a newly introduced pose similarity feature space (PSFS) obtained from extra-pose (different pose) and intra-pose (same pose) similarities in the training phase. This, as will be explained shortly, effectively avoids the problem of re-training the classifier for each of the class separately when a new pose class is added. For a test image it outputs a measure of confidence or probability for all poses without explicitly estimating underlying densities.

To show the effectiveness of our method, in the presence of large illumination differences and expression variations, our system is evaluated on CMU-PIE face

database. It is chosen since it is the largest database that exhibits each subject in large pose and illumination differences. Also, it enables us to compare our results with some of the recent pose estimation methods that uses the same database and are able to recognize poses in the presence of large lighting variations.

6.1.2 Database setup

We used a subset of PIE database to evaluate our pose estimation algorithm. The portion of PIE database, we used, consists of 21 illumination differences of 68 subjects at 13 poses. Each 640x480 pixel PIE image is converted to greyscale. A 128x128, closely cropped face part is used. We note that this is standard procedure and any state of the art face detector like [RBK98] [JV03] can be used for this purpose. A subject imaged under 21 illumination variations (with background lighting off) is shown in Figure 6-1. Out of 13, the main 9 poses are considered for estimation whereas the remaining 4 poses corresponding to up/down tilt (pitch) of the face (see Figure 6-2) are treated as previously unseen test poses, in order to see how well the system performs in assigning these into one of the corresponding poses.

Figure 6-1: A subject from PIE imaged under 21 illumination conditions.

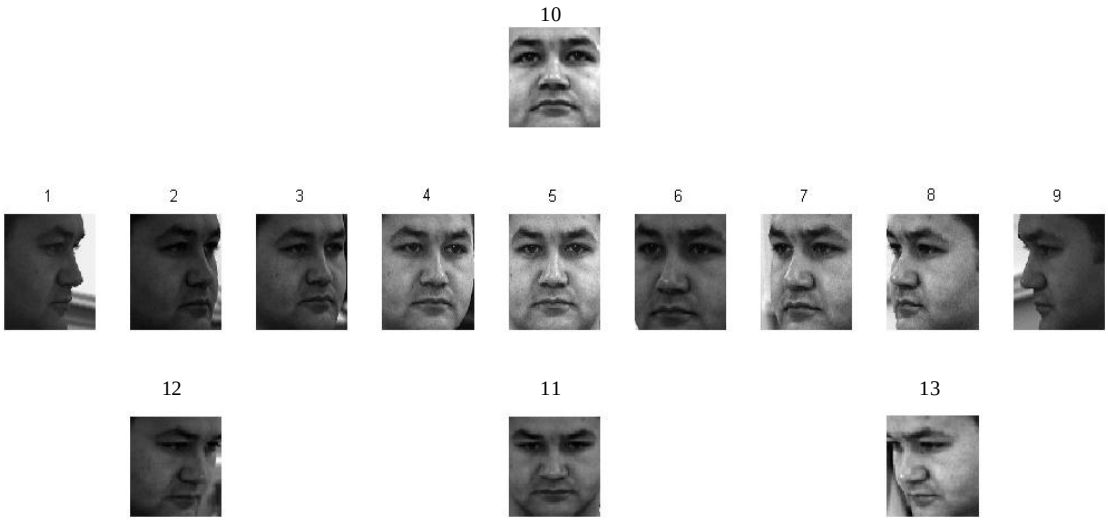


Figure 6-2: The main 9 pose variations in PIE along with 4 pitch variations in the corresponding poses.

Each pose is approximately 22.5^o apart with full right profile +90^o (pose 1), frontal 0^o (pose 5) to left profile -90^o (pose 9). 15 subjects are used for training and rest of 53 subjects for testing.

For training, the main 9 poses (pose 1 -9) in 4 different illumination variations (out of 21) and 3 PIE expression variations per subject in each pose are considered, see figure 6-3 for an example. Following the PIE naming convention illumination variations correspond to flash 01, 04, 13 and 16, which capture well the extent of illumination variations present (see figure 6-1). The expression variations are neutral, smiling and blinking at frontal lighting.

For testing, all 21 illumination conditions per subject are considered, where the 17 illumination conditions, not used in training, provides a way to assess the performance of the pose estimation method in previously unseen lighting variations.

In such scenarios one can expect that there will be a huge overlap between nearest or adjacent poses in the derived feature space. We therefore introduce a new classification framework that overcomes this and models well the in-pose variations due to large illumination and expression changes.

Figure 6-3: (Along Rows) All 9 pose variations in CMU-PIE; pose 1(right profile) to pose 9 (left profile) views; (Along columns) 7 imaging conditions; illumination and expression variations.

6.2 Proposed Approach

For the reasons stated earlier, we solve the pose estimation as a classification problem from a machine learning point of view. Instead of directly modelling the extracted LESH features and solve it as a multiclass problem, we rather use similarity scores of these features within same pose and among different poses. This implies the construction of a new feature space based on these computed similarities. Such an approach has huge benefit in that it effectively turns a multiclass problem into a binary two-class one while still representing well all the in-pose variations. We model this new feature space.

6.2.1 Pose Similarity Feature Space 'PSFS'

We transform the whole problem into a new feature space termed as pose similarity feature space (PSFS). This PSFS is derived by computing similarities between LESH features, coming from same pose examples and similarities between features from all the different pose examples.

As measure of similarity, we use modified Kullback-Leibler (KL) divergence [KL51], also known as Jeffrey-divergence, which is numerically stable, symmetric and robust with respect to noise and size of histogram bins [RTG00]. It actually gives a measure of dissimilarity between two histograms. Thus low values means more similar. It is defined as

$$d(H,K) = \sum_r \eta_r \sum_i \left(h_{i,r} \log \frac{h_{i,r}}{m_{i,r}} + k_{i,r} \log \frac{k_{i,r}}{m_{i,r}} \right) \tag{6.1}$$

where, subscript 'r' runs over total number of regions(partitions) and 'i' over number of bins in each corresponding local histogram h and k, 'm' is the corresponding bin's mean and 'η_r' is used as a provision to weigh each region of the face while computing similarity scores. This could be used, for instance, in overcoming the problems due to expressions, by assigning a lower weight to regions that are mostly affected. In our experiments, for now, this η is set to 1.

For each example in our training set, we compute these similarities with the rest of the examples in the same pose on derived LESH features. Concatenating them, give rise to an intra pose 'IP' (same pose) similarity vector. Similarly computing these similarities for each example with all other examples in a different pose give rise to an extra pose 'EP' similarity vector. Thus each example is now represented in a PSFS as a function of its similarities by these IP or EP vectors.

Note however, the dimensionality of this PSFS is a direct function of the total number of examples per pose in the training set. Therefore to put upper limit on the dimensionality of this derived PSFS and also to generate many representative IP and EP vectors for a test face, as explained shortly, we partition our training sets into some disjoint subsets in such a way that each subset has same number of subjects in each pose.

To understand it better, consider, for example, our training set comprising of 15 subjects, where each subject is in 7 different illumination and expression imaging conditions in each of the 9 poses, see figure 6-3. Therefore we have 15x7(105) examples per pose.

Deriving a PSFS directly means a 105 dimensional feature space, while partitioning it into some disjoint subsets, such as each subset has all the 15 subjects but in some different combination of the imaging conditions, would yield a 15 dimensional features space while still representing all the variations we want to model.

6.2.2 Formal description of our approach

Formally, our approach is that we first partition the training set into 'k' disjoint subsets (all N training examples per pose per subset), the subsets are disjoint in terms of the 7 imaging conditions (chosen such as each subject is at a different imaging condition in that subset).

In each subset, we then compute for each example, its similarity to the rest of the examples in the same pose on derived LESH features. Thus for 'N' examples per pose, we compute 'N-1' similarities for each example, concatenating them, give rise to a 'N-1' dimensional intra-pose (IP) similarity feature vector for each of the N examples. Extra-pose (EP) vectors are obtained similarly by computing these similarities between each example in one pose with N-1 examples in a different pose by leaving the same subject each time (since we use a symmetric similarity measure). Thus we will have $(N \times P \times K)$ IP samples and $K\left(N\sum_{i=1}^{P-1}(P-i)\right)$ EP samples for training. Where 'N' is number of examples/pose and 'P' is the total number of pose.

Although there will be a large number of EP samples as compared to IP in the derived PSFS but we note that, IP samples tend to have low values as compared to EP and form a compact cluster in some sub-space of the PSFS. This is validated in Figure 6-4 which shows a 3-D scatter plot of IP and EP samples from one of the subset, by randomly choosing 3 features from IP and EP similarity vectors. Note that IP samples are depicted from all of the 9 poses while only those EP samples are depicted which are computed among large pose variations, such as between frontal and left/right profile view or between left and right profile view. The scatter plot is shown in logarithmic

scale for better viewing. Figure 6-4 provides an intuitive look at how the problem is easily separable when there are large pose variations, while EP samples coming from nearest pose examples can be seen as causing a marginal overlap with the IP class.

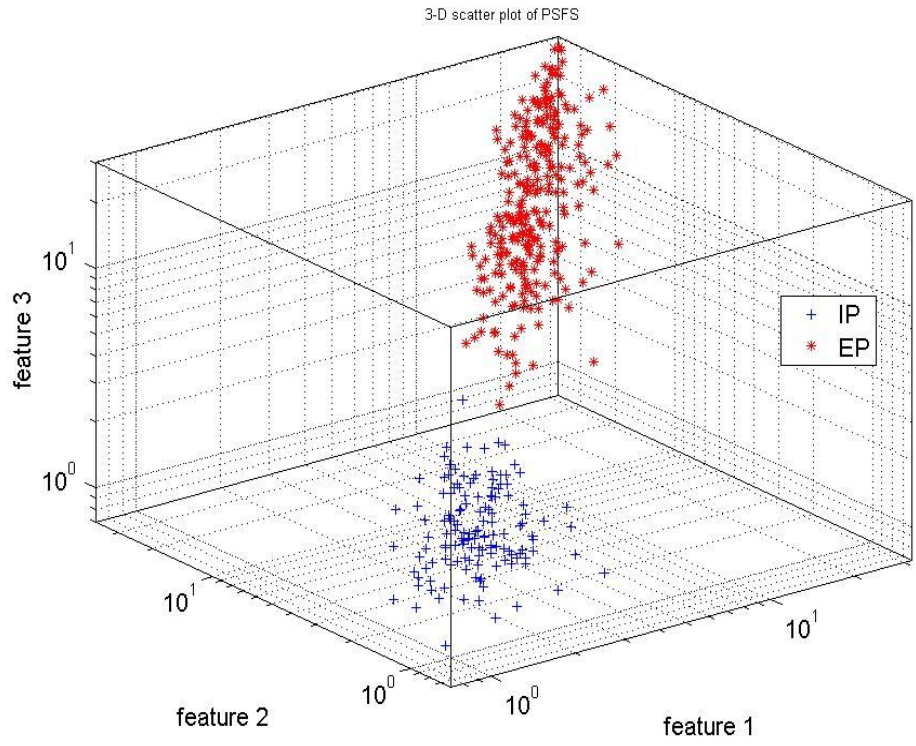


Figure 6-4: 3-D scatter plot of IP and EP vectors from one of the subset. IP samples are drawn by randomly choosing 3 features from IP vectors from all of the 9 poses, while EP samples are depicted only for large pose variations i.e. between frontal and left or right profile or between left and right profile view.

The training set is used as a gallery and thus for a test face, computing its similarity with all of the examples in each pose in each subset of the gallery produces many representative similarity vectors for that test image. Therefore there is a good chance that more of the similarity vectors, coming from where the pose of the test face and gallery are same, falls in the IP class as compared to those which are coming from even slight pose variations.

To learn the structure of this PSFS, we therefore seek to separate the two classes. Any classifier can be used for this purpose. However, since the way problem is posed, a classifier that can find a nonlinear boundary between the two classes is preferred. We therefore use a simple AdaBoost classifier originally proposed by Schapire and Singer [SS99]. AdaBoost is known for its good generalization ability and an extremely good performance in binary classification tasks. An AdaBoost classifier using nearest neighbour rule in each iteration is trained in this feature space for that purpose. It provides a non-linear boundary between the two classes.

For a test image, k vectors are obtained for each pose by computing similarities from N-1 subjects in each pose in each training subset. All of these are classified to belong to either of the class. Final decision is then made by considering only those classified as IP, and assigning the label of the pose from which majority of these are coming. This probability for each pose is calculated simply by:

$$\gamma_p = \frac{n_p}{K} \tag{6.2}$$

where n_p is the number of IP vectors computed from the corresponding gallery pose p and K is the total number of IP vectors possible for each pose p, that always corresponds to the number of subsets. It is then further normalised such as probabilities for all poses always sum to one, the final probability for each pose is therefore computed as

$$\mathrm{P}(\gamma_p) = \frac{\gamma_p}{\sum_{p=1}^{P} \gamma_P} \tag{6.3}$$

As stated earlier, the rational of making subsets of training set is now evident, as on one hand it limits the dimensionality of the feature space, while still representing well all the in-pose variations, and on the other hand it generates many representative vectors per pose for a test image, which provides us with a probability score and helps in overcoming the short comings of the classifier itself.

6.3 Experimental Setup and Results

As described in the preceding section, for the 15 training subjects we have (105)15x7 examples per pose. We partition them into 7 disjoint sets (each with 15 examples) for each pose as described earlier. This generates a 14 dimensional PSFS by computing all the IP and EP vectors using LESH features. AdaBoost is then trained on this PSFS.

For a test face, after extracting LESH feature vector, we compute similarities with 14 examples in each pose, for each training subset. This will generate one 14 dimensional similarity vector for each representative pose in each subset; therefore, we will have 7x1x9 (63) similarity vectors for that particular test image. They are then classified as either IP or EP. Those which are assigned label as IP are then further used to compute probability scores, by using Equations 6.2 and 6.3, for each of the 9 poses.

Final pose estimate is based on by assigning the pose, which has the highest score. This way we hope to overcome the problem of any misclassified nearest pose EP vectors.

For the 53 test subjects, we considered three sets of tests to evaluate our pose estimation method. In the first, we use test images at seen imaging conditions and poses. In the second, we use test images at unseen 17 illumination conditions, and in the third, we evaluate how well the system does in assigning the 4 unseen poses (pitch variations) to the corresponding 1 of the 9 poses.

6.3.1 Test results for seen imaging conditions

In this test, the input images are at one of the imaging conditions and poses seen in the training stage. The test set, therefore, consists of 4 illumination differences and 3

expression variations of the 53 subjects (15-68) not included in the training set at each of the nine poses. Thus, there are 53x7x9 (3339) test images.

Each test image is classified as one of the pose based on the probability scores obtained from Equation 6.3, this corresponds to rank-1 rates. Whereas in order to see how well it performs in assigning a given image to the nearest poses, the results are summarized in the form of a confusion matrix. Figure 6-5 provides average estimation results for each pose. While Table 6-1 summarizes the classification results obtained on all the (3339) test examples in a confusion matrix. The overall average estimation accuracy is 84.06% in terms of rank-1 rates and 96.62% for estimates within $\pm 22.5^{o}$ of accuracy.

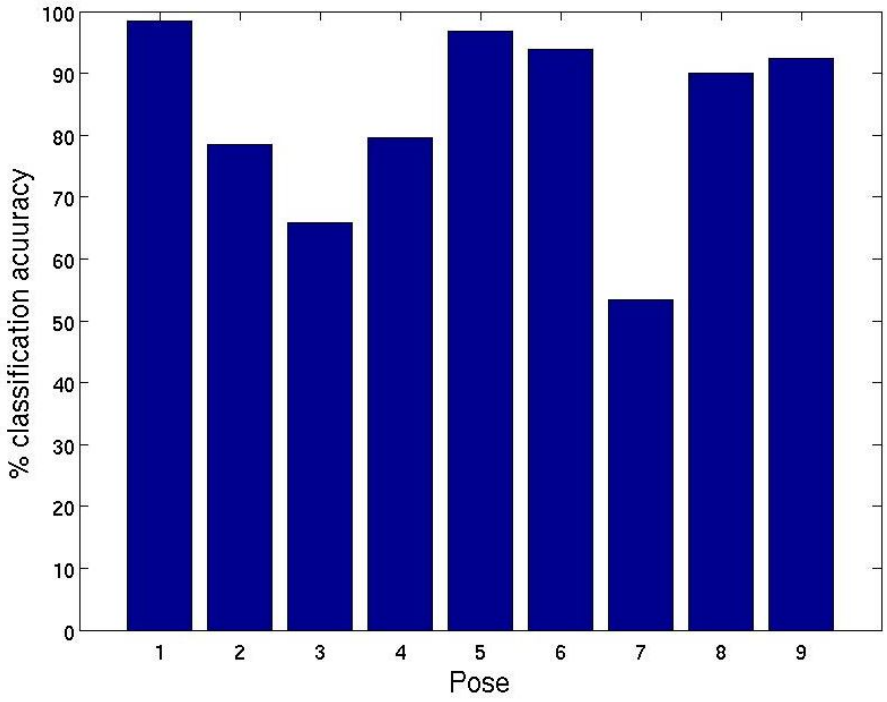


Figure 6-5: Average classification scores for each pose

Table 6-1: Confusion Matrix for test examples at seen imaging conditions

		System Pose Estimates								
		1	2	3	4	5	6	7	8	9
TRUE POSE	1	365	4	2	0	0	0	0	0	0
	2	47	291	24	6	2	0	0	1	0
	3	31	59	244	31	5	1	0	0	0
	4	0	0	21	295	39	2	0	0	0
	5	0	0	0	10	359	2	0	0	0
	6	0	0	0	0	17	348	5	1	0
	7	0	0	1	0	23	44	198	85	20
	8	0	0	0	0	0	4	11	334	22
	9	0	0	0	0	0	2	4	22	343

In confusion matrix, the rows entries are indexed by the true pose of the input images, while column entries are labelled by our classification procedure-determine pose. The entries on the diagonal indicate the number of correctly classified images at each pose. The sum of each row is 371 (an entry of 371 on the diagonal indicates perfect classification for that pose).

6.3.2 Test results for previously unseen illumination conditions

In this test, we evaluate our pose estimation system on images of the 53 test subjects that exhibit one of those illumination differences not used in training. The test set, therefore consists 17 illumination differences of 53 subjects, all at neutral expression, at each of the nine poses. This amounts to 53x17x9 (8109) test examples.

The results are reported similarly in a confusion matrix. An entry of 53x17 (901) on the diagonal indicates perfect classification for that pose. An average estimation accuracy of 85.15% is achieved in terms of rank-1 rates and 97.5% for estimates within $\pm 22.5^{o}$ of accuracy.

Table 6-2: Confusion Matrix for test examples at 17 unseen illumination conditions

		System Pose Estimates								
		1	2	3	4	5	6	7	8	9
TRUE POSE	1	845	44	12	0	0	0	0	0	0
	2	58	774	68	1	0	0	0	0	0
	3	21	42	745	71	22	0	0	0	0
	4	0	0	8	749	139	5	0	0	0
	5	0	0	0	21	868	12	0	0	0
	6	0	0	0	13	141	701	31	15	0
	7	0	0	0	2	34	105	643	85	32
	8	0	0	0	0	0	27	108	711	55
	9	0	0	0	0	1	2	4	25	869

6.3.3 Test results for unseen poses

The four poses 10, 11, 12 and 13 that corresponds to the pitch variations of the corresponding poses 5, 2 and 8 respectively are used in this test to evaluate the performance of the system for unseen poses. The system should estimate the pose of the test images at one of these four unseen poses as one of the corresponding adjacent pose in the training set, e.g. an image at pose 10 should be assigned a pose label of 5 or 6. Therefore, here the performance of the system should primarily be assessed with regards to estimating the pose within $\pm 22.5^{o}$ of accuracy (adjacent poses).

The test set in this case consists of images of 53 subjects at each of the four unseen poses with all 21 lighting differences. It, therefore, amounts to 53x21x4 (4452) test examples. An entry of 53x21 (1113) at the corresponding pose in a row indicates perfect classification. The average overall rank-1 accuracy in this case is 63.1%, whereas the performance of system with-in $\pm 22.5^{o}$ of accuracy is 97.1 %.

Table 6-3: Confusion Matrix for 04 unseen poses at 21 illumination conditions

		System Pose Estimates								
		1	2	3	4	5	6	7	8	9
True Pose	10	0	0	19	111	**791**	168	24	0	0
	11	0	0	16	153	**684**	241	19	0	0
	12	95	**701**	307	9	1	0	0	0	0
	13	0	0	8	0	0	41	267	**637**	168

6.4 Discussion and Conclusion

We can compare our results with few of the recent works [YC05] and [PC05] which use same database and approximately the same setup, where former achieved 82.4% rank-1 and later achieved 84.11% rank-1 and 96.44% within $+22.5^{o}$, they however, pre-registered a test face to top 3 to 4 poses by using 3 landmark locations on the face and did not include expression variations.

The recognition at previously unseen illumination conditions is slightly better than the seen ones. This is surprising, but it may be due to the fact that the effect of shadow at most of these illumination differences is much less than those of the extreme illumination variations included in the training set. The extracted LESH features are sensitive to this since a strong casting shadow may introduce an unwanted edge and hence affects the energy distribution in the corresponding local region. Our system achieves best recognition scores most of the time on full profile views, the reason, perhaps, stems from the fact that a face at these views is most distinguishable in terms of pure shape. Since our system is build on a pure shape representation, these results provides an intuitive relation if one looks at the corresponding cropped faces at these poses, see Figure 6-2.

The performance of our method on registering a given face to the nearest pose (adjacent poses) is above 97%. It provides us with probabilities for each pose and that makes it very attractive from a practical stand point, since this can be used directly as our confidence in a given pose in the further face recognition stage.

On concluding remarks, we have presented a front-end pose estimation system which functions in presence of large illumination and expression changes. A new feature description that encodes the underlying shape well is proposed, and an efficient classification procedure is suggested which turns the multi-class problem into a binary one and solves the problem of discriminating between nearest poses.

Based on proposed LESH feature description, we introduced to generate a generic similarity feature space, that not only provides an effective way of dimensionality reduction but also provides us with many representative vectors for a given test feature vector. This is used in generating probability scores for each pose without explicitly estimating the underlying densities, which is very useful in later face recognition across pose scenarios.

We hope that the proposed feature description and the notion of modelling the similarity space will prove very useful in similar computer vision problems.

PART-III

Probabilistic Models for Pose Invariant Face Recognition

Chapter 7. Statistical Models for Automatic Pose Invariant Face Recognition

In this chapter we specifically address the problem of recognition across pose when there is only one image per person available for training. Recognizing a person reliably having seen only one image and from previously unseen view point is a very challenging problem. Current state-of-the-art methods address this problem by explicitly learning the appearance variations of a face at different views. This can be achieved by using a pool of generic faces at different views and attempting to model how a face relates at two different views. Here we assume that only one image per person (e.g. frontal) is available in the database. The relationship between frontal and non frontal images is treated as a statistical learning problem.

7.1 Introduction

Several previous studies have presented algorithms which can take a single probe image at one pose and attempt to match it to a single gallery image at a different pose. One successful approach is to create a full 3D head model for the subject based on just one image [RBV02] [BGPV05] and compare the 3D models. This is the current-state-of-the-art for view independent face recognition from a single image. A drawback of this, however, is a precise registration of the probe in order to guide the fitting process and moreover the computation involved is too restrictive for a practical face recognition system.

An alternative approach in the 2D context is to treat this as a learning problem in which we aim to predict frontal images from non-frontal ones. The other approach in this

context is to only model what is discriminative between images of the same subject in different poses and images of the different subjects in these poses. Lucey and Chen [LC08] categorized these into view-point generative and view-point discriminative approaches respectively. The emphasis in view-point generative methods is to find a mapping function that can be used to generate a given non-frontal image to its frontal counterpart [BP95] [ZC00] [GMB04] [BGPV05]. A simple distance metric is then utilized to compute similarity between a frontal gallery image and the transformed probe image. The assumption in generative methods, of finding a transformation function that can be used to generate a near perfect frontal image of its non-frontal counterpart in any pose, is not realistic and in practice the transformed images exhibit strong variations that accounts for degradation in recognition performance. The view-point discriminative approaches, on the other hand, has some inherent advantages over the viewpoint-generative approach as more emphasis is given to discrimination, rather than the generation of a gallery view image from the probe view appearance [KY03] [LC05] [KK05] [LC08]. This, however, is a very naive assumption in that it assumes that the appearance variations among different subjects across different pose mismatches is always larger than variations among same subject in different poses. This practically may not be true since the discrimination of appearance across large pose differences of the same subject becomes significant enough that it can not preserve the identity.

We can overcome these associated problems by first finding a generative function for each pose and then following a view-point discriminative approach to model the associated appearance variations specific to each pose explicitly. The goal of such an approach is creating a model that can predict how a given face will appear when viewed at different poses. This seems a natural formulation for a recognition task especially in unconstrained scenarios. In this chapter we develop this idea in a full Bayesian probabilistic setting.

7.1.1 Overview

In the context of the recent discriminative or generative appearance based methods the emphasis has been to directly model the local appearance change, due to pose, across same subjects and among different subjects. Differences exist among different methods

in how these models are built, but the goal of all is same i.e. trying to approximate the joint probability of a gallery and probe face across different pose. Such an approach is particularly attractive in that it automatically solves the one training image problem in a principled way as these appearance models can be learned effectively from an offline database of representative faces, see Figure 7-1. Another benefit of such a line of work is that adding a new person's image in the database does not require training the models again. We note, however, almost all of these methods proposed in literature until now intrinsically assume a perfect alignment between a gallery and probe face in each pose. This alignment is needed, because, otherwise, in current appearance-based methods it is not possible to discern between the change of appearance due to pose and change of appearance due to the local movement of facial parts across pose.

In this chapter, we introduce novel methods in this line of work and propose to build models on features which are robust against misalignments and thus do not require the facial landmarks to be detected as such. Our approach, briefly, is to learn statistical models describing the approximated joint probability distribution of a gallery and probe image at different poses. Since we address the problem where at most one training image (e.g. frontal) is available, we learn such models by explicitly modeling facial appearance change between frontal and other views when identity of a person is same and when it is different across pose. This is done by computing similarities between extracted features of faces at frontal and all other views. The distribution of these similarities is then used to obtain the likelihood functions of the form

$$P(I_g, I_p \mid C) \quad , \quad C \in \{S, D\} \tag{7.1}$$

where C refers to classes when the gallery I_g and probe I_p images are similar S and dissimilar D in terms of subject identity. For this purpose an independent generic set of faces, at views we want to model, is used for offline training. Figure 7-1 illustrates the underlying concept.

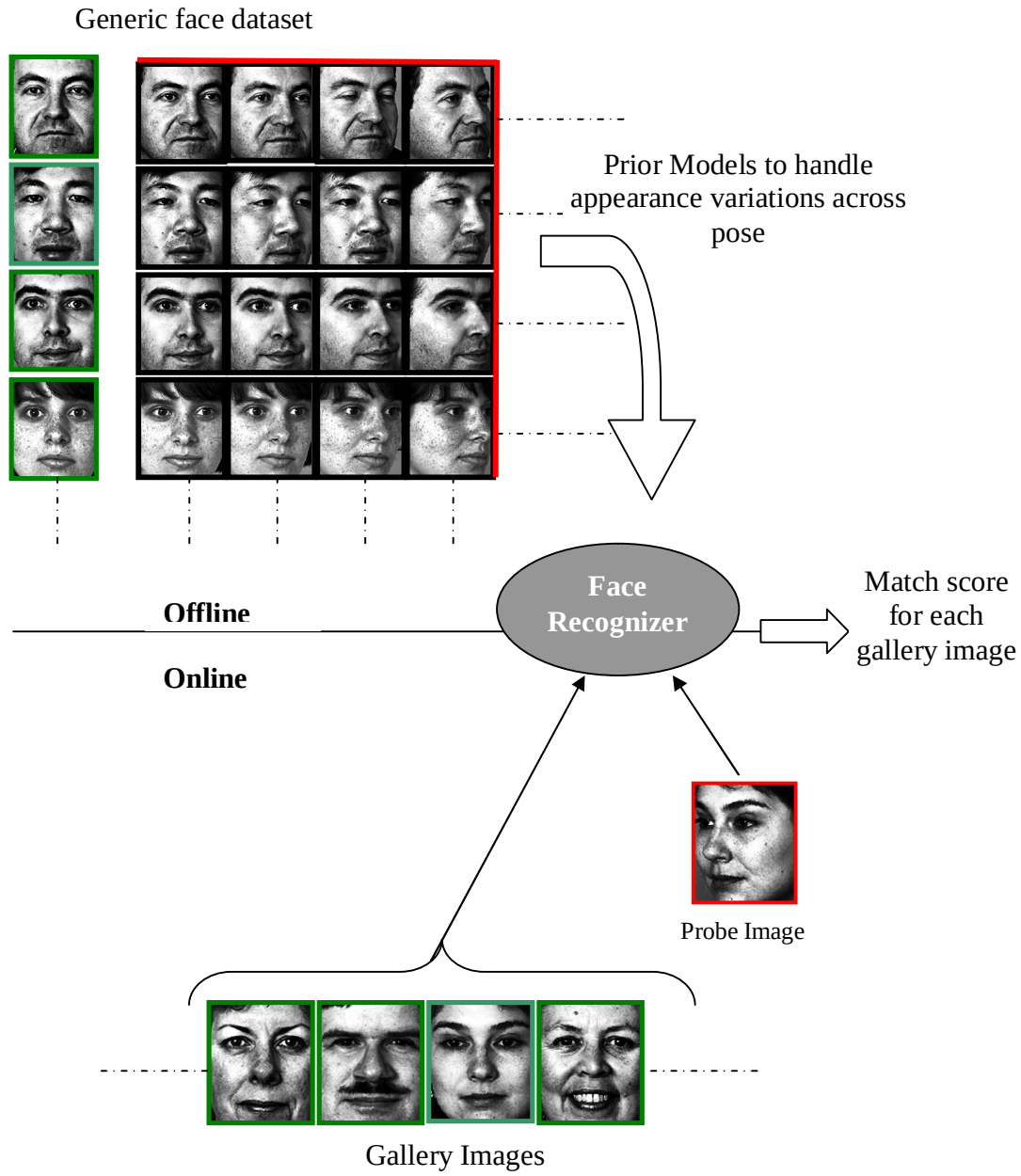


Figure 7-1: Recognition using single image. The offline component trains the face recognizer on how to handle mismatch. Offline images are representative of the appearance variations we anticipate seeing in the gallery and probe images.

A contribution is made towards improved recognition performance across pose without the need of properly aligning gallery and probe images. To achieve this, we propose to use face-GLOH-signature description introduced in chapter 3. This feature description captures the whole appearance of a face in a rotation and scale invariant manner, and is shown robust with regards to variation of facial appearance due to misalignments in chapter 4. Furthermore, we propose to synthesize these features at non-frontal views to frontal by using multivariate regression techniques. The benefit of this in recognition performance is demonstrated empirically. To approximate the likelihood functions in Equation 7.1, local kernel density estimation as opposed to commonly used Gaussian model is suggested for deriving these prior models.

7.1.2 Related work

Our contribution lies in the body of work that concerns estimating the joint likelihood for the purpose of recognition in the presence of pose mismatch. Here, therefore, we introduce the related existing work in this direction in order to put our work into the right context.

There are three main methodologies in this domain. The first tries to model the joint likelihood function $P(I_g, I_p \mid S)$ directly in the presence of pose mismatch. The likelihood $P(I_g, I_p \mid D)$ is typically omitted due to the complexity associated with its estimation. Due to the large dimensionality of the whole face I, subspace methods based on PCA are employed to approximate the likelihood from a generic face dataset. To match a known gallery image I_g against an unknown probe I_p the following approximation is then used

$$I_g \cong \int_I I \cdot P(I, I_p) dI \tag{7.2}$$

The vectors I_g and I_p are then matched using some canonical distance measure. The Tensorface [VT02] and Eigen Light Field [GMB04] are very recent techniques that fall into this category.

The second methodology attempts to model the differential appearance between gallery and probe images. In order to make the approximation, an offline generic set of examples is used.

$$P(I_g, I_p \mid C) \cong P(\left|I_g - I_p\right| \mid C) \tag{7.3}$$

These likelihoods are attempting to model the whole face, for both similar and dissimilar classes by using absolute difference in pixels. The most well known method in using differential appearance has been the intra-personal and extra-personal approach of Moghaddam and Pentland [MP97]. The similar and dissimilar classes of the differential appearance likelihood in equation 7.3 are modeled through a normal distribution. These distributions are estimated within a subspace, found using PCA. Techniques centered on LDA [KK05] also employ a similar paradigm, in terms of differential appearance, although they are not framed within a strict probabilistic framework.

The third methodology is to decompose the face into an ensemble of salient patches/regions. [BP93] [M02] reported superior recognition performance, in the presence of occlusion and expression variations, with respect to approaches that treat the face as a whole. Recently Kanade and Yamada [KY03] proposed an effective technique for pose invariant recognition within this framework. Their extension was centered on the hypothesis that individual patches/regions can be treated as independent and modeling the change of appearance of these small regions is more effective as opposed to the whole face appearance. They thus approximate the appearance likelihood as follows.

$$P(I_g, I_p \mid C) \cong \prod_{i=1}^{R} P(s_i \mid \theta_{C_i}) \tag{7.4}$$

where s_i is the sum of squared difference (SSD) in pixels between gallery and probe patch r at position i, and θ is the assumed Gaussian parametric form. Their approach, therefore, can be thought of a direct extension of the differential appearance paradigm in that they combine the differential appearance likelihoods of several local patches to approximate the joint likelihood. Some extensions to this approach have been reported in the literature [LC05] [LC08] with improved performance.

We argue, however, that the strong assumption of patch independence is not statistically right, since face is a highly symmetrical object and different regions of a

face are not independent. Such an assumption, nonetheless, was needed in order to overcome the problems that arise, in modeling differential change of appearance across pose, in holistic appearance based methods. In this thesis, we suggest that one should, instead, derive such whole-face appearance representations that are easily tractable across pose and can take into account, to certain extent, the change of appearance of different parts of face due to pose.

As mentioned previously, a commonality exists among all these approaches in that they try to build models based on pixel appearances and thus require a perfect alignment between gallery and probe images. This alignment is usually artificially imposed in the face normalization stage by using at least 3 or more facial landmark points, such that all of these points are in a fixed location in the final image. We note that detecting several facial landmarks with pixel accuracy is not trivial and thus localization of face in different poses for the purpose of registration can never be guaranteed. For the purpose of automatic pose invariant face recognition, this is a major bottleneck for above mentioned methods.

7.2 Modeling Whole-Face Appearance Change across Pose

Our approach is to extract whole appearance of the face in a manner that is robust against misalignment due to localization. For this we use face-GLOH-signature feature description introduced earlier. It models the local parts of the face and combines them into a global description. Figure 7-2 shows example feature extraction for two subjects at different poses from PIE database.

As noted earlier, following a view-point discriminative approach directly by modeling the appearance across each pose can not preserve the identity in general, across large pose differences, especially when we do not assume a strict alignment between images. We therefore, synthesize the obtained feature vectors at non-frontal views to frontal by finding a generative function for each pose and then follow a view-point discriminative approach to model the remaining associated appearance variations

specific to each pose explicitly. Computing similarities using these synthesized features between frontal and other poses provides us with prior distribution for each pose.

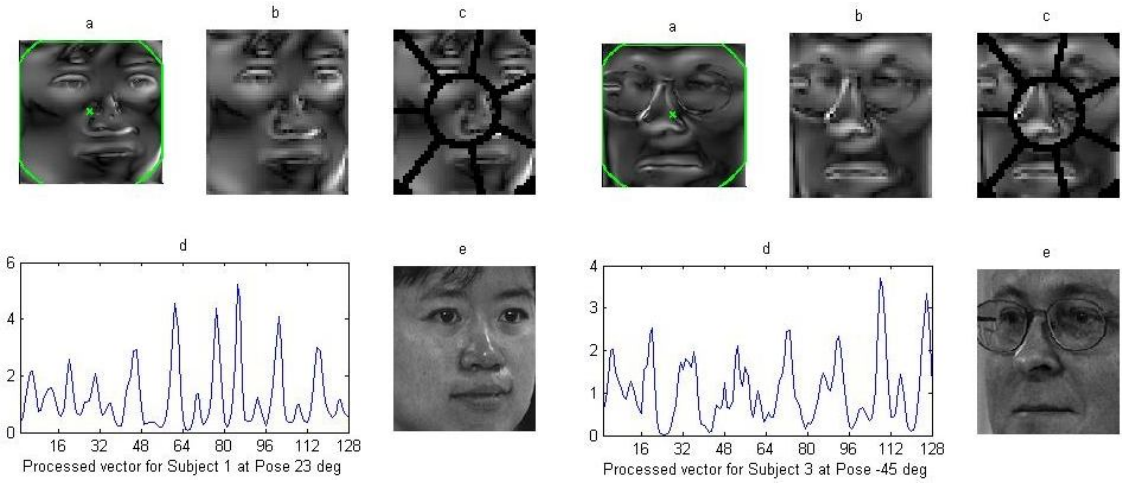


Figure 7-2: Face-GLOH-Signature extraction of two subjects at different poses.

7.2.1 Generative pose models for synthesizing features

It is well known that when a large number of subjects are considered, the recognition performance of appearance-based methods deteriorates significantly. It is due to the fact that distribution of face patterns is no longer convex as assumed by linear models. By transforming the image into a scale and rotation invariant manner, we assume that there exists a certain relation between these features of frontal and posed image that we can linearly transform. We justify this assumption by comparing the similarity distributions estimated from non-synthesized features and synthesized features. One simple and powerful way of relating these features is to use the regression techniques. Let us suppose that we have the following multivariate linear regression model, for finding relation between the feature vectors of offline gallery I_F (frontal) and any other probe I_P view examples.

$$\mathbf{I}_F = \mathbf{I}_P \cdot \mathbf{B}$$

$$\begin{bmatrix} \vec{I}_{F1}^T \\ \vdots \\ \vec{I}_{Fn}^T \end{bmatrix} = \begin{bmatrix} \vec{I}_{P1}^T & 1 \\ \vdots & \\ \vec{I}_{Pn}^T & 1 \end{bmatrix} \begin{bmatrix} \beta_{(1,1)} & \cdots & \beta_{(1,D)} \\ \vdots & \ddots & \vdots \\ \beta_{(D+1,1)} & \cdots & \beta_{(D+1,D)} \end{bmatrix} \tag{7.5}$$

where n corresponds to number of examples such that $n > D+1$, with D being the dimensionality of each $\vec{I}_F$ and $\vec{I}_P$. Note that $\vec{I}_F$ and $\vec{I}_P$ are column vectors corresponding to each other in terms of subject identity. **B** is a pose transformation matrix of unknown regression parameters. Under the sum-of-least-squares regression criterion, **B** can be found using Moor-Penrose inverse.

$$\mathbf{B} = (I_P^T I_P)^{-1} I_P^T I_F \tag{7.6}$$

This transformation matrix **B** is found for each of the poses I_P (e.g. $\pm 22.5^\circ, \pm 45^\circ, \pm 65^\circ, \pm 90^\circ$) with frontal 0° I_F. Given a set of *a priori* feature vectors representing faces at frontal I_F and other poses I_P, we can thus find the relation between them.

Any incoming probe feature vector can now be transformed to its frontal counterpart online using:

$$\widehat{I_P} = \mathbf{B}_P \cdot [I_P^T \quad 1]^T \tag{7.7}$$

7.3 Obtaining Prior Appearance Models for Recognition

The likelihood of joint occurrence of a probe and gallery face at different poses is obtained by using an offline generic set of faces at views we want to model. These models explicitly describe the appearance variations between frontal and other poses for when the identity is same and when it is different. These prior models can be used to compute a match score between an online probe and gallery image in a Bayesian setting.

We approximate the joint likelihood of a probe and gallery face as:

$$P(I_g, I_p \mid C, \phi_g, \phi_p) \approx P(\gamma_{pg} \mid C, \phi_g, \phi_p) \tag{7.8}$$

where $C \in \{S, D\}$ refers to classes when the gallery I_g and probe I_p images are similar (S) and dissimilar (D) in terms of subject identity. ϕ is the pose angle for the corresponding gallery and probe image and γ_{pg} is the similarity between gallery and probe image. These likelihoods for the similar and dissimilar class are then found by modeling the distribution of similarities of extracted features between frontal and every pose from offline training set. Cosine distance is used as a similarity metric

$$\gamma(I_g, I_p) = \frac{I_g \cdot I_p}{\|I_g\|_2 \cdot \|I_p\|_2} \tag{7.9}$$

where $\|\cdot\|$ denotes the Euclidean norm of the vectors.

Figure 7-4 depicts the histograms for the prior same and different distributions of the similarity γ for gallery and probe images across a number of pose mismatches. These distributions depicted here are obtained by using images of half of the subjects in all 9 poses from PIE database. To make the estimation of pose transformation matrix **B** feasible, 200 images in each pose of these subjects (illumination and expression variants) are used. The images used in our evaluation are cropped from the database without using any commonly employed normalization procedure. The face images therefore contain typical variations that may arise from miss localization such as back ground, part clippings and scale. Example images at frontal and 4 pose differences of a subject from PIE database are shown in figure 7-3.

Figure 7-3: Example images of a subject at 5 poses

Note that, in Figure 7-4 the more separated the two distributions are the more discriminative power it has to tell if the two faces are of the same person or not for that particular pose. It is clear that the discriminative power decreases as the pose moves away from frontal. As shown, synthesizing features to frontal significantly improves this discrimination ability over a wide range of poses.

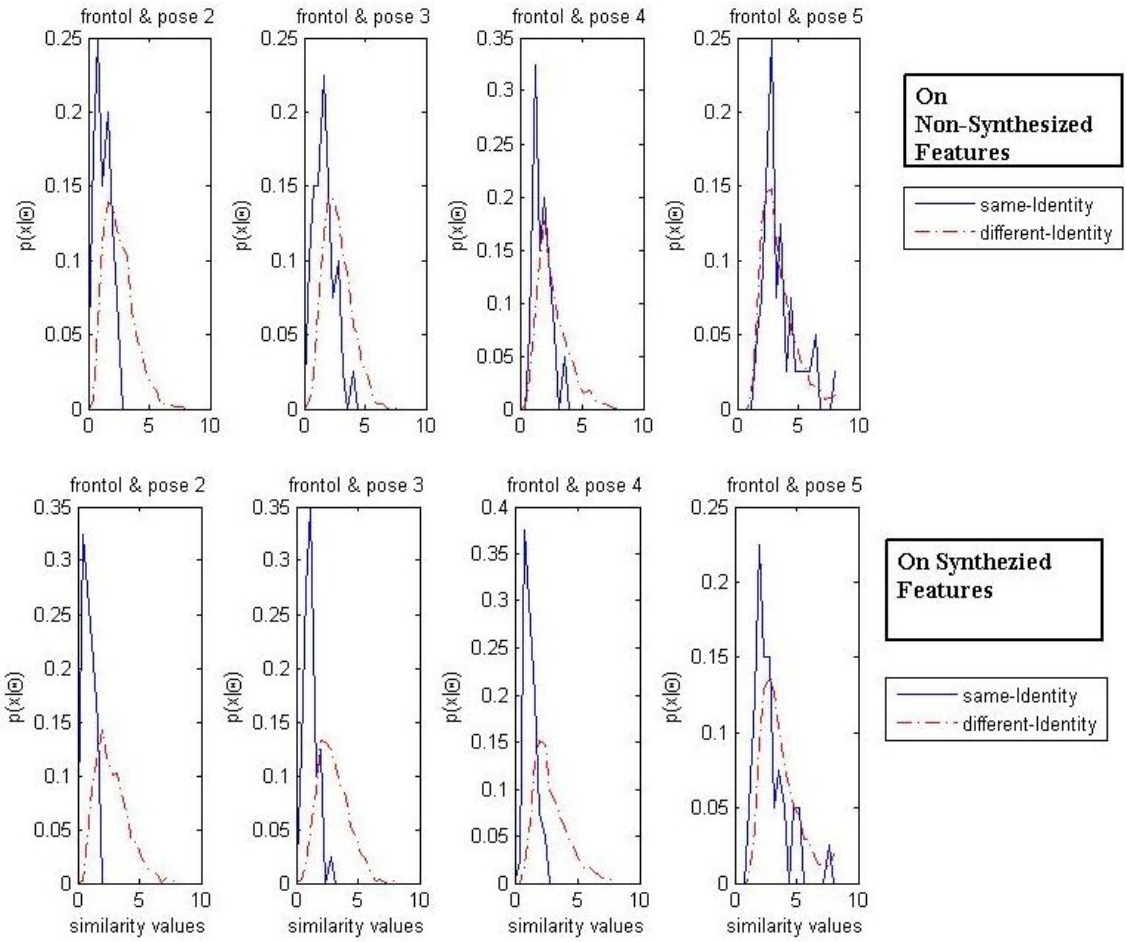


Figure 7-4: x-axis denotes the similarity measure γ and y-axis denotes the density approximation. First row depicts histograms for the same and different classes on non-synthesized features across 4 pose mismatches (see figure 7-3 for the approximate pose angles). Second row depicts the kind of separation and improvement we get by using feature synthesis.

7.3.1 Local kernel density estimation

In order to compute $P(\gamma_{pg} \mid S, \phi_p)$ and $P(\gamma_{pg} \mid D, \phi_p)$[1] i.e. conditional probabilities describing similarity distributions when subject identity is same S and when it is different D, these distributions must be described by some form. The most common assumption is the Gaussian. Functional density estimators like the Gaussian assume a functional form of the distribution and therefore depend heavily on the accuracy of that assumption. It is especially deceitful that a functional estimation always 'looks' correct, no matter how poor the assumption is for the underlying distribution. In [LC06] authors noted that describing such prior appearance models by a normal density is not optimal and results in biased recognition results. We also note that employing a normal density results in a poor fit, see Figure 7-5. We therefore propose to use local kernel density estimate.

$$P(\gamma) = \frac{1}{N\sigma^n} \sum_{i=1}^{N} k\left(\frac{\gamma - \gamma_i}{\sigma}\right) \tag{7.10}$$

where,

$$k(\nu) = \frac{1}{(2\pi)^n} e^{\frac{\nu^2}{2}} \tag{7.11}$$

There exist various methods to automatically estimate appropriate values for the width σ of the kernel function [W07]. In this work, we simply set σ to be the average nearest neighbor distance:

$$\sigma^2 = \frac{1}{N \sum_{i=1}^{N} \min_{i \neq j} \left| \gamma_i - \gamma_j \right|^2} \tag{7.12}$$

As depicted in Figure 7-5, the kernel density estimate is a better fit. It is because the assumption of Gaussian distribution in such scenarios is generally not fulfilled. Kernel density estimator, on the other hand, is known to approximate arbitrary distributions [S92].

[1] Note that angle ϕ_g is typically omitted since the gallery pose is fixed to frontal.

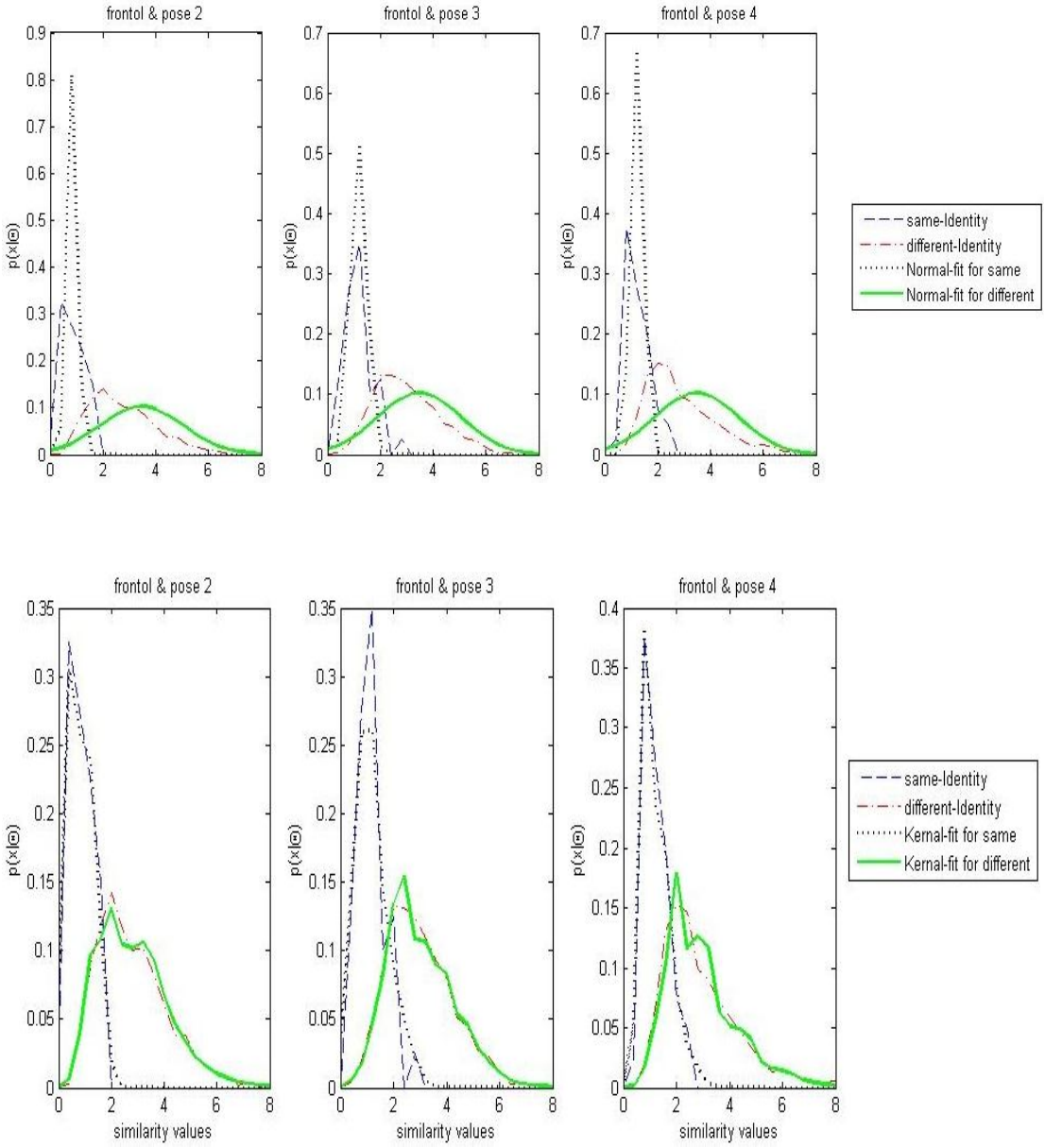


Figure 7-5: 1st row shows fitting a normal density, 2nd row shows the kernel density fits on the distribution of similarities obtained previously.

7.4 Recognition across Pose

Obtained likelihood estimates $P(\gamma_{pg} \mid S, \phi_p)$ and $P(\gamma_{pg} \mid D, \phi_p)$ in the previous section, can now be directly used to compute the posterior probability.

For an incoming probe image I_P at some pose ϕ_p, of unknown identity, we can now decide if it is coming from the same subject as gallery I_g, with each of the gallery image, by using this posterior as a match score. Employing these likelihoods, using Bayes rule, we write:

$$P(S \mid \gamma_{pg}, \phi_p) = \frac{P(\gamma_{pg} \mid S, \phi_p)P(S)}{P(\gamma_{pg} \mid S, \phi_p)P(S) + P(\gamma_{pg} \mid D, \phi_p)P(D)} \tag{7.13}$$

Since the pose ϕ_p of the probe image is in general not known, we can marginalize over it. In this case the conditional densities for similarity value γ_{pg} can be written as

$$P(\gamma_{pg} \mid S) = \sum_p P(\phi_p)P(\gamma_{pg} \mid S, \phi_p) \tag{7.14}$$

and

$$P(\gamma_{pg} \mid D) = \sum_p P(\phi_p)P(\gamma_{pg} \mid D, \phi_p) \tag{7.15}$$

Similar to the posterior defined in equation 7.13, we can compute the probability of the unknown probe image coming from the same subject (given similarity γ_{pg}) as

$$P(S \mid \gamma_{pg}) = \frac{P(\gamma_{pg} \mid S)P(S)}{P(\gamma_{pg} \mid S)P(S) + P(\gamma_{pg} \mid D)P(D)} \tag{7.16}$$

If no other knowledge about the probe pose is given, one can assume the pose prior $P(\phi_p)$ to be uniformly distributed. We, however, use the pose estimates for a given probe face by our developed front-end pose estimation procedure as detailed in chapter 5 and 6. The pose estimation system provides us with probability scores for each pose that can be used directly as priors $P(\phi_p)$ in Equation 7.14 and Equation 7.15. Due to a reasonably high accuracy of these pose estimates, these probabilities can act as very strong priors and thus increase the chances of a probe to be recognized correctly.

We compute this posterior for an unknown probe image with all of the gallery images and choose the identity of the gallery image with the highest score as recognition result. Figure 7-6 details the general flow of the developed framework in a block diagram.

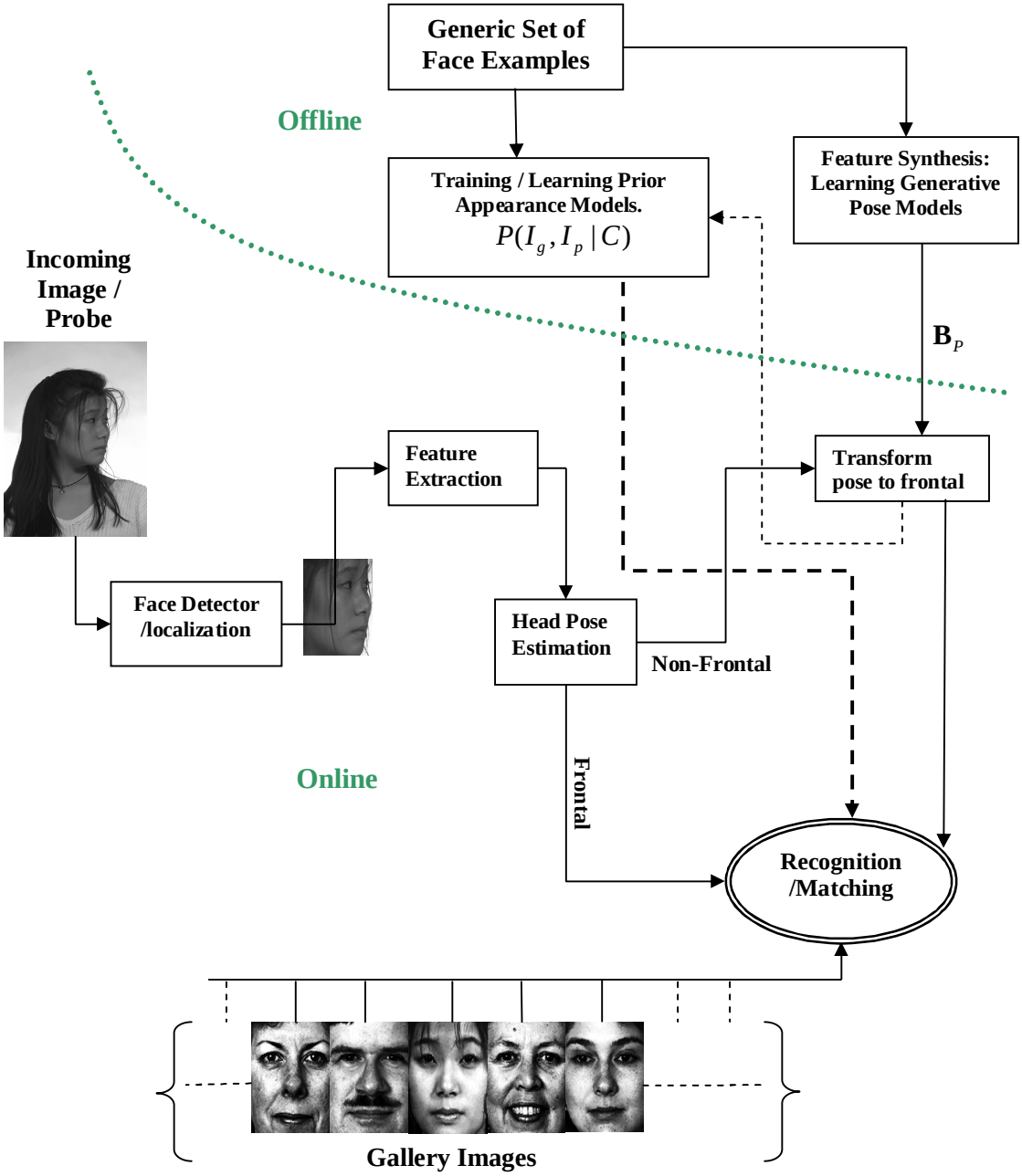


Figure 7-6: Overview of the developed fully automatic face recognition system

7.5 Conclusion

We have presented a pose invariant face recognition method that requires only a single image of the person to be recognized in the gallery. The proposed approach is centered on modeling joint appearance of gallery and probe images across pose in a Bayesian framework. We have proposed novel methods in this direction by introducing to use a more robust feature description as opposed to pixel-based appearances. The variation of these features across pose is modeled by a multivariate regression approach. Furthermore using kernel density estimate, instead of commonly used normal density assumption, is proposed to derive the prior models. Our method does not require any strict alignment between gallery and probe images and that makes it particularly attractive as compared to the existing state of the art methods.

In the next chapter we present several experimental results and comparisons with the previous state-of-the-art methods in order to show the effectiveness of the proposed method.

Chapter 8. Automatic Pose Invariant Face Recognition Results

In this chapter we test the method and models developed in the previous chapter on two of the large databases i.e. CMU-PIE database and FERET database. Where Images of half of the subjects from PIE and FERET are used as generic offline set for training of the models, while remaining are used for gallery and probe sets.

8.1 Experimental Setup

The pose subsets of both the databases are used in our experiments. From CMU-PIE all images of 68 subjects imaged under 13 poses in neutral expression are used. The approximate pose difference between images of the same subject is 22.5^o varying from frontal 0^o to $\pm 90^o$ profile. FERET set includes images of 200 subjects in 9 pose variations with an approximate pose difference of 15^o, varying from 0^o frontal to $\pm 60^o$ profile view.

We test our system for the automatic case without any manual localization. We, therefore, use Viola-Jones face detector [JV03] to automatically localize faces. Detected Face windows are then cropped from the database without employing any commonly used normalization procedure. Therefore images contain typical variations that may arise due to miss-localization like scale, part clippings and background. All images are then resized to 128x128 pixels. Typical variations present in the database are depicted in few of the example images in Figure 8-2. Note that, since we do not employ any kind of normalization such as detecting and fixing eye location or eye-distance, face images across pose suffer from typical misalignments.

Figure 8-1: Examples of detected face windows depicting typical variations due to misalignments e.g. scale, part clipping, background etc.

As mentioned, half of the subjects are used for training the models, this amount to images of 32 subjects for the experiments on PIE database and 100 subjects for the experiments on FERET.

As the gallery, the frontal images of all the subjects are used. For the tests on PIE database, since we do not assume any alignment between gallery and probe images, therefore models are trained for the main 9 poses i.e. pose 1-9 in Figure 8-2. While pose 10,11,12 and 13 corresponding to up/down tilt of the face are treated as the variations due to misalignment for corresponding poses in the test set. All 13 poses for a subject in the test set are therefore considered. Class priors $\mathrm{P}(\mathrm{S})$ and $\mathrm{P}(\mathrm{D})$ are set to $P(S) \ll 1$ and $P(D) = 1 - P(S)$ in all of our experiments.

8.2 Test Results

We provide results of several experiments demonstrating the effectiveness of our method. The first set of results are obtained by using PIE database

Figure 8-2: 13 poses covering left profile (9), frontal (1) to right profile(5), and slightly up/down tilt in pose 10, 11, 13 and 12 of corresponding poses in 8, 1 and 4 respectively.

8.2.1 Experiment 1: Known probe pose

For our first experiment, we assume probe pose to be known and therefore use equation 7.13 to compute the posterior. In order to show the effectiveness of our method, we include the results of Kanade and Yamada's [KY03] Bayesian Face Sub-region 'BFS' algorithm and Eigenface algorithm [MP97] for comparison, where the former is considered since our method is similar to it in the principled approach while Eigenface is included as it is the common benchmark in facial image processing. Results are reported on PIE database.

In order to use BFS on our dataset, the face image is divided into 32x32 pixel overlapping patches, with an overlap of 16 pixels. This is done since we do not assume aligned probe and gallery with respect to eyes and mouth positions. Prior models are obtained as described in the original paper [KY03] by using sum of square difference measure for each patch. Our results, as depicted in Figure 8-3, shows the robustness of our method against misalignments between probe and gallery while the results of BFS are much worse on misaligned images as compared to what they originally reported on the same database, see Table 8-1.

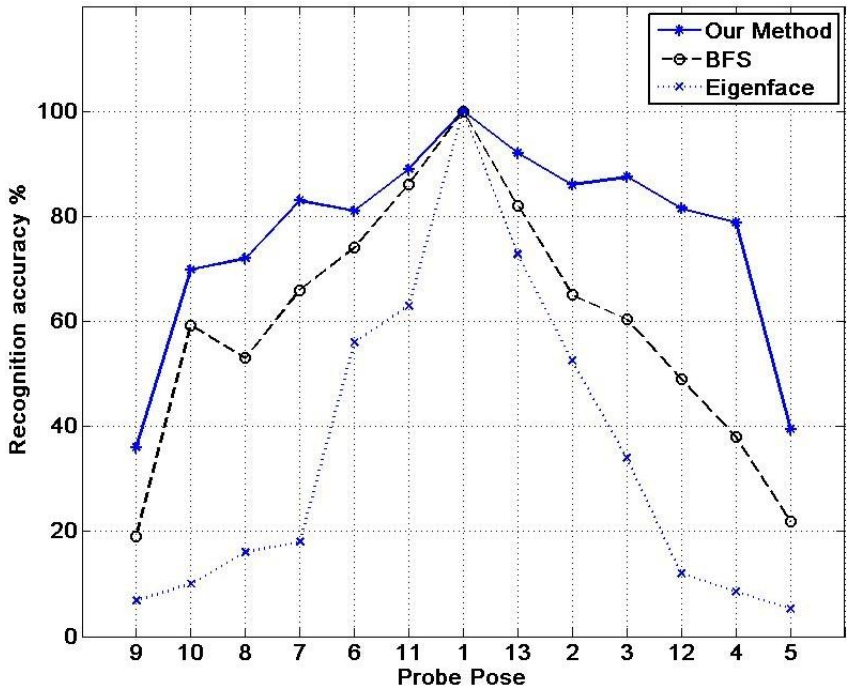


Figure 8-3: Recognition performance for each of the 13 PIE poses for the test set. Results of our method and comparison with BFS and Eigenface for known probe pose.

8.2.2 Experiment 2: Unknown probe pose

For our second experiment, we report results for the fully automatic case, where pose of the probe images is not known *a priori*.

For an incoming probe image, we extract face-GLOH-Signature as described in chapter 3. In order to synthesize these features to frontal we need to know the probe pose, as we have to use the corresponding pose transformation matrix $\mathbf{B}$. Since we use the front-end pose estimation step, as described earlier that provides us with the probabilities for different possible poses, we therefore use marginalization (Equation 7.14 and 7.15) by transforming the extracted feature vector of given probe to frontal for all poses. Note that, still using marginalization after a pose estimation step may seem counterproductive at first but since our system is based on models learned from

synthesized features, and as shown in Figure 7-4, the distributions depicting the similar class are almost same for the nearest pose mismatches, therefore this in fact improves the recognition performance in most cases. This is due to the fact that these pose prior probabilities, obtained from the pose estimation system, act as weights and they are only high for the nearest poses. As shown in Figure 8-4 performance of our method with marginalization, by using strong pose priors, improves recognition accuracy.

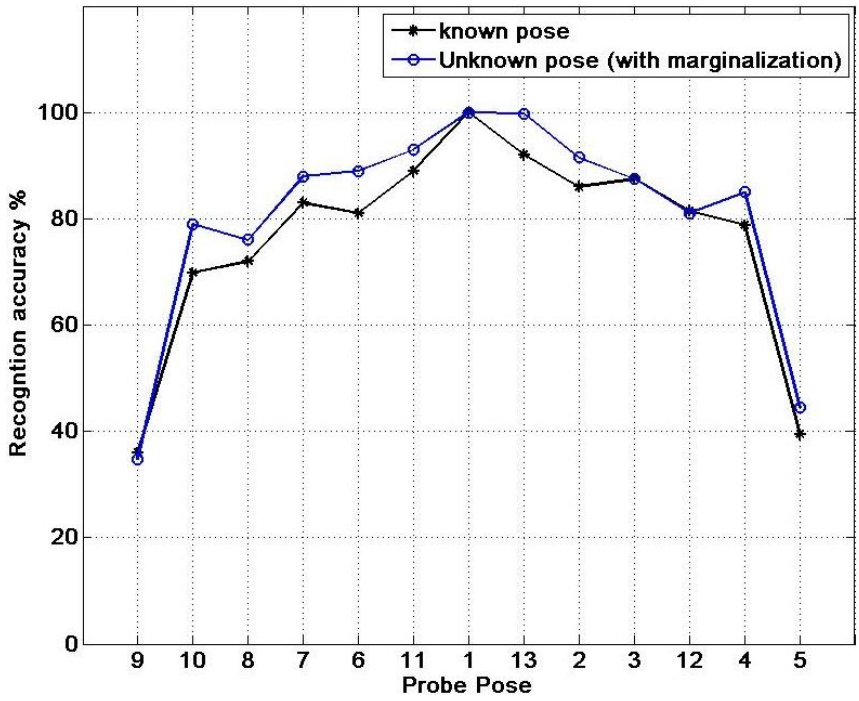


Figure 8-4: Comparison of recognition performance with and without marginalization on PIE database.

8.2.3 Experiment 3: Comparison with and without feature synthesis

We compare the performance with and without feature synthesis. Figure 8-5 shows the performance gain achieved by using feature synthesis. As much as 20% of performance

gain is observed for probe poses moving away from frontal. These results provide an insight into the effectiveness of the learned generative pose models.

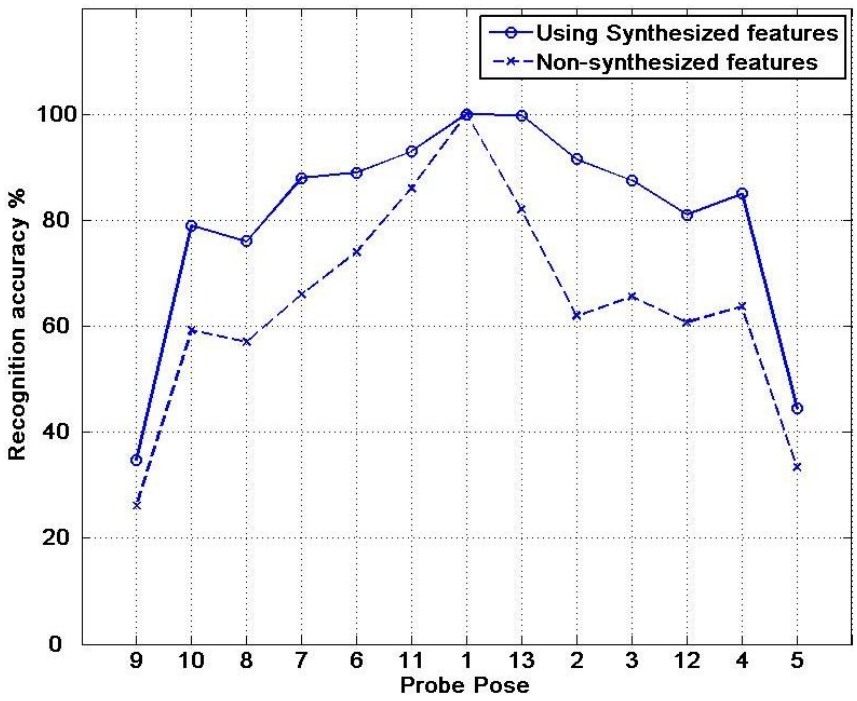


Figure 8-5: Comparison of our method for with and without feature synthesis (results reported here are obtained using marginalization).

8.2.4 Experiment 4: Evaluation on FERET

Following a similar procedure, the results on FERET are summarized here. We use 100 subjects as probe where frontal images of all 200 subjects are used as gallery. Note that there is a significant pose difference between FERET and PIE images, the average pose difference across FERET is 15^{o}. The results reported here are obtained in a fully automatic setting where probe poses are obtained by using marginalization. The recognition accuracy up till $\pm 45^{o}$ of pose difference is above 90%. The overall average recognition accuracy on FERET is 92.1%. Figure 8-6 plots the average recognition accuracy across each probe pose.

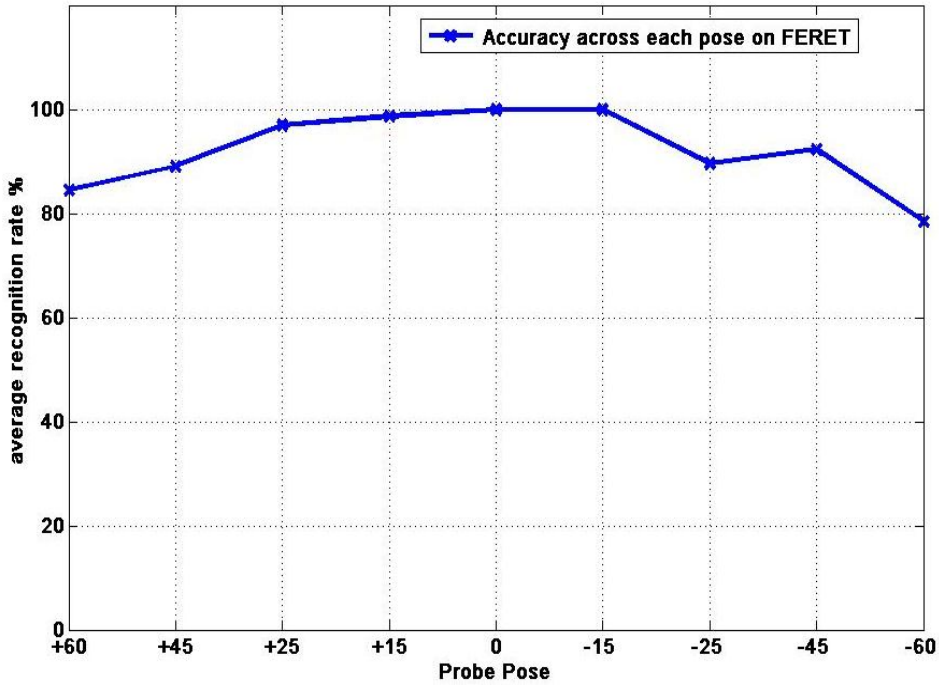


Figure 8-6: Average recognition accuracy across each pose on FERET

8.2.5 Experiment 5: Recognition across databases

For all experiments shown so far the training and gallery/probe subjects were taken from the same database. Here we demonstrate the generalization ability of the learned models across different databases. For this purpose prior models are learned by using FERET database and tested on PIE database. Note, however, there is a significant pose difference among the two databases. In order to cope with that we use 7 PIE poses (pose 1, 2, 3, 4, 6, 7 and 8) that loosely correspond to the corresponding FERET poses ($0^o, \pm 25^o; \pm 45^o, \pm 60^o$). In Figure 8-7 we show recognition accuracies for tests with seven poses of the PIE database. Training for both prior appearance models and pose generative model is performed by using all the 200 subjects of the FERET database. Correspondence between FERET and PIE poses is determined manually.

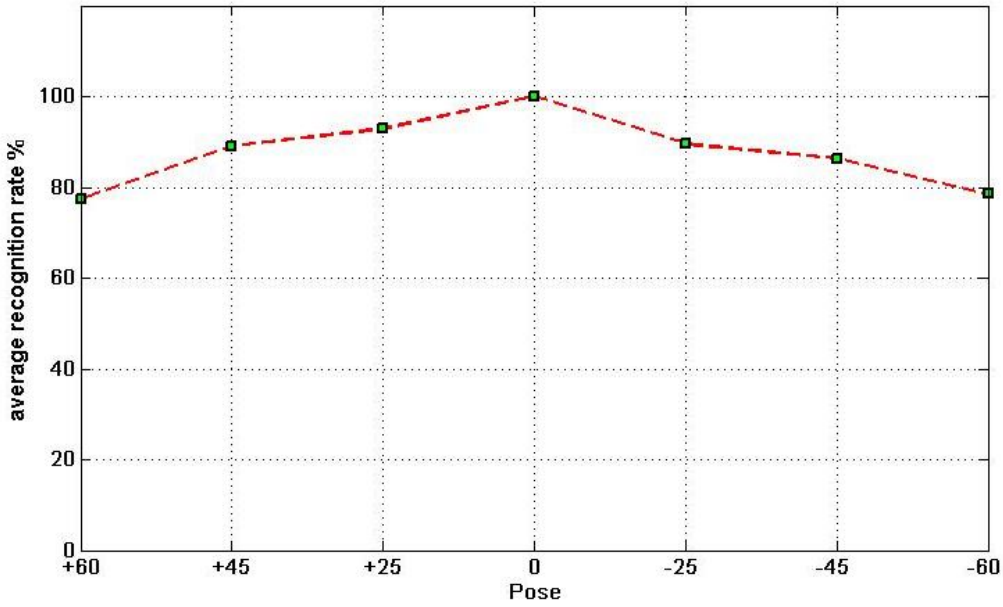


Figure 8-7: Recognition accuracy for test on 7 PIE poses. Prior models are obtained using all the FERET subjects and tested using only the PIE subjects.

The results indicate the good generalization ability of our method. The overall accuracy is 87.7% that compares favorably with the results obtained previously using the same respective database for training and testing.

8.3 Comparison with Contemporary Methods

We summarize and compare the average recognition accuracy of our method with that of some of the most representative algorithms that achieved state-of-the art results on face identification studies on the same databases so far. When comparing identification results, one should keep in mind the over restrictive assumptions behind these methods that hinders a direct generalization of these methods to fully automatic case. In particular the degree of manual intervention should be noted. Almost all of the previous studies rely on manual localization of some points on the face in order to align images or to establish a direct correspondence to model local patches around these points. Our

method does not require any manual registration and is fully automatic in this sense. With these considerations in mind, we present a summary of identification performance from other studies in Table 8-1.

Table 8-1: Comparison with state-of-the-art face identification studies across pose

Study	Alignment	Database	Pose Difference	% Correct
Kanade and Yamada [KY03]	3 points	CMU-PIE	Average on all 13 poses	81
Gross et al. [GMB04]	3 points	FERET	Average on all 9 poses	75
Gross et al.	39 points	CMU-PIE	Average on all 13 poses	78.8
Blanz et al. [BGPV05]	11 points	FRVT	45^o	86
Chai et al. [CCG07]	3 points	CMU-PIE	23^o / 45^o	98.5 / 89.7
Prince et al. [PEWF08]	21 points	CMU-PIE	45^o / 67.5^o	100 / 91
		FERET	15^o / 45^o	100 / 99
Our Method	No alignment	CMU-PIE	Average on all 13 poses 23^o / 45^o / 67.5^o	80.4 91.5 / 87.9 /81
Our Method	No alignment	FERET	Average on all 9 poses 15^o / 45^o / 60^o	92.1 100 / 90.3 / 82.5

Our results compares favorably with the previous approaches. Gross et al. (Eigen light filed method) [GMB04] reports an overall 75 percent first-match results over 100 subjects from the FERET database, by using three manually marked feature points. Our system achieves 92.1 percent performance, with out manual registration. In the same study, they also report 39 percent and 93 percent performance for the PIE database conditions 67.5^o and 23^o, respectively, with a large number (> 39) of manually labeled key points. For the same conditions, we report 81 percent and 91.5 percent, respectively, with no annotation. Kanade and Yamada's BFS method achieves an average of approximately 81% performance on PIE database. Their method however, is sensitive to the manually annotated points on the face. As has been shown in our experiments (Figure 8-3) the performance of their method is much worse when we do not assume any

alignment. Blanz et al. [BGPV05] report results for a test database of 87 subjects with a horizontal pose variation of 45^o from the Face Recognition Vendor Test FRVT 2002 [FRVT02] database, using, on the average, 11 manually established feature points. They investigate estimating the 3D model and creating a frontal image to compare to the test database (86.25 percent correct). Our system produces better performance at larger pose differences for comparable databases.

Probably the best results reported up till now are those of Chai et al. [CCG07] and Prince et al. [PEWF08]. Both of the methods use a similar pose contingent linear transformation of non-frontal views to frontal, where Chai et al. synthesize raw pixels and thus it cannot generalize directly to the automatic case where typical variations due to miss localization are expected, while Prince et al. transforms the local features extracted from manually located 21 points on the face image (at a different pose) to frontal and then model the variation of the corresponding local features across pose, their method therefore puts a hard constraint on the precise correspondence of these points across each pose. The comparison in Table 8-1 shows that our method is able to achieve comparable or better results in a fully automatic sense even without the need of properly aligning the gallery and probe images. That is especially attractive in the context of fully automatic pose invariant face recognition.

8.4 Conclusion

We have presented several experimental results and comparisons with previous state-of-the-art approaches, demonstrating the effectiveness and weakness of the proposed approach. Our method is able to achieve above 80 % of performance within a pose difference of approximately 65^o.

The performance of our system, as depicted by experiments on PIE database, on full profile views i.e. 90^o of pose difference is around 45 %. A relatively low performance on these conditions depends on a number of factors and among others suggests that using a linear model for pose transformation is not able to cope well with these extreme pose differences.

Note that when a probe is at frontal, the scores are 100% since exactly the same images are used in the gallery for frontal pose.

The results reported here are for the fully automatic case, where faces are localized by using a face detector and we do not assume probe pose to be known. Our results show that one can achieve comparable performance without requiring the facial landmarks to be detected. Current methods rely on this information for the purpose of registration. For the purpose of automatic pose invariant face recognition, this is a major bottleneck for the current methods. Our approach tries to lift off this barrier and works directly on the output of a face detector. Although, we have presented results by using gallery as fixed at frontal pose, we note that it is straight forward to use our method for any pose as gallery.

Chapter 9. Outlook and Future Directions

This dissertation is concerned with the automatic machine recognition of human faces for the purpose of identification in unconstrained scenarios. Here we present a cursory overview of the main contributions and some directions for future work.

9.1 Summary

We have elaborated on the background and specifics of an automatic face recognition process. More specifically, we differentiate the face localization from alignment and argue that the alignment stage has to be avoided in order to realize a fully automatic system especially when considering the pose variations. A brief but comprehensive literature review of the current state of the art methods is presented.

A comprehensive account of almost all the feature extraction methods used in current face recognition systems is presented. Specifically we have made distinction in the holistic and local feature extraction and differentiate them qualitatively as opposed to quantitatively. It is argued that a global feature representation should be preferred over a bag-of-features approach. The problems in current feature extraction techniques and there reliance on a strict alignment is discussed. We have introduced to use face-GLOH signatures that are invariant with respect to scale, translation and rotation and therefore do not require properly aligned images. The resulting dimensionality of the vector is also low as compared to other commonly used local features such as Gabor, LBP etc. and therefore statistical based methods can also benefit from it.

In a typical multi-view face recognition task, where it is assumed to have several examples of a subject available for training, we have shown in an extensive experimental setting the advantages and weaknesses of commonly used classification methods. Our

results show that under more realistic assumptions, most of the classifiers failed on conventional features. While using the introduced face-GLOH-signature representation is relatively less affected by large in-class variations. This has been demonstrated by providing a fair performance comparison of several classifiers under more practical conditions such as misalignments, large number of subjects and large pose variations. An important conclusion is to be drawn from the results is that conventional multi-view face recognition can not cope well with regards to large pose variations. Even using a large number of training examples in different poses for a subject does not suffice for a satisfactory recognition

Large pose differences cause significant appearance variations that in general are larger than the appearance variation due to identity. One possible way of addressing this is to learn these variations across each pose, more specifically by fixing the pose and establishing a correspondence on how a person's appearance changes under this pose, one could reduce the in-class appearance variation significantly. The pose information is valuable and for a fully automatic face recognition system the pose of the incoming probe/test image has to be known in this context. An effective pose estimation framework for the purpose of automatic face recognition is developed. The proposed front-end pose estimation system functions in the presence of large illumination and expression changes. In this context we have introduced a new feature description based on local energy model of feature perception, termed as LESH, which encodes the underlying shape and is insensitive to skin color and illumination variations. Based on proposed LESH feature description, we introduced to generate a generic similarity feature space, that not only provides an effective way of dimensionality reduction but also provides us with many representative vectors for a given test feature vector. This is used in generating probability scores for each pose without explicitly estimating the underlying densities, which is very useful in later face recognition across pose scenarios.

Finally we integrate the head pose estimation system with a novel fully automatic face recognition system. We introduced a pose invariant face recognition method that requires only single image of the person to be recognized in the gallery. The proposed approach is centered on modeling joint appearance of gallery and probe images across pose in a Bayesian framework. We have proposed novel methods in this direction by

introducing to use a hybrid approach to generative and discriminative models. The generative pose models are obtained by modeling the variation of the suggested features across pose by a multivariate regression approach. Furthermore using kernel density estimate, instead of commonly used normal density assumption, is proposed to derive the discriminative prior appearance models. The method provides us with a full posterior over possible gallery matches which can also be easily used for face authentication. Our method does not require any strict alignment between gallery and probe images and that makes it particularly attractive as compared to the existing state of the art methods. We have presented several experimental results and comparisons with previous state-of-the-art approaches, demonstrating the effectiveness and weakness of the proposed approach.

9.2 Future work

The different methods introduced and developed in this dissertation may be extended in many ways. In unconstrained scenarios, one of the main problems addressed in this regard is the misalignment or poor localization of face caused by a face detector stage. The problem of misalignments is addressed by using a whole face representation that is robust against variations caused by scale, rotation or translation of the facial parts. Our approach however is sensitive to the detection output. In particular the noise caused by face detector module such as back ground, face part clippings etc. induce unwanted appearance variations that compromise the recognition accuracy. This can be improved, by using a refinement stage e.g. using procedures to precisely localize only the face in the detected support of the image. To automate this, an effective approach would be to use human skin color. Skin colors are distributed over a small area in the chrominance plane with the major difference between skin tones being variations in intensity [MM99]. After detecting a skin probability image an ellipse fitting method can be used to approximate the skin blob. The work in [BS03] details one such procedure.

The performance of our system, on full profile views i.e. 90^o of pose difference is around 45 %. A relatively low performance on these conditions depends on a number of factors and among others suggests that using a linear model for pose transformation is

not able to cope well with these extreme pose differences. In this regard a non-linear approach to model the underlying generative process can be followed. The non-linear regression can be carried out by projecting the data into kernel space, see [B06] for recent developments in this direction.

In order to address the single training image problem, we have proposed to follow a hybrid approach by explicitly modeling the joint distribution of a gallery and probe image across pose. Our framework uses a very simple Bayesian formulation. More complex Bayesian models can be used to approximate the true underlying generative distributions. Moreover our current formulation only incorporates a single probe image, in many situations such as video feeds we can have several images of a same subject to be recognized. In such cases the proposed framework can be easily extended to use multiple probe images and hence improve performance.

With the substantial increase in processing powers and computational speeds of the digital computers, it is perhaps best to use 3D methods along with 2D approaches. Future works on addressing face recognition may benefit from using multi modal 2D+3D approach, promising works in this directions are [BGPV05] [ANRS07].

Bibliography

[A01] W. Atkins. A testing time for face recognition technology. Biometric Technology Today, 9(3):8–11, 2001.

[AHP04] T. Ahonen, A. Hadid, and M. Pietikainen, “Face recognition with local binary patterns,” ECCV, pp. 469–481, 2004.

[ALC08] Ashraf A.B., Lucey S., and Chen T. Learning patch correspondences for improved viewpoint invariant face recognition. In IEEE CVPR, June 2008.

[ANRS07] A. F. Abate and M. Nappi and D. Riccio and G. Sabatino,” 2D and 3D face recognition: A survey” Pattern Recognition Letters. Volume 28(14), 2007.

[ANSI04] ANSI INCITS 385-2004 “Information technology - Face Recognition Format for Data Interchange”.

[B06] Bishop, C. M.” Pattern recognition and machine learning”. New York: Springer. 2006.

[BA00] Beumier, C. & Acheroy, M.. Automatic 3D face authentication. Image and Vision Computing, 18, 315-321. 2000.

[BA00] Baudat, G. & Anouar, F. E., Generalized discriminant analysis using a kernel approach. Neural Computation, 12, 2385-2404, 2000.

[B96] Beymer D.: Pose-invariant face recognition using real and virtual Views. M.I.T., A.I. Technical Report No.1574, March 1996.

[BCF06] K. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition. Computer Vision and Image Understanding, 101(1):1–15, 2006.

[BGPV05] V. Blanz, P. Grother, P. Phillips, and T. Vetter. Face recognition based on frontal views generated from non-frontal images. In Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition CVPR, Volume 2, pages 454–461, 2005.

[BHK97] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection”, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, pp. 711-720, 1997.

[BMS02] Bartlett, M. S., Movellan, J. R., & Sejnowski, T. J. (2002). Face recognition byindependent component analysis. IEEE Transactions on Neural Networks, 13, 1450-1464.

[BP93] Brunelli R. and Poggio T.: Face recognition: Features versus templates. IEEE Trans. PAMI, vol. 15, no. 10, pp. 1042–1052, 1993.

[BP95] Beymer, D. Poggio, T., Face recognition from one model view. In International conference on computer vision, 1995.

[BP96] D. Beymer and T. Poggio, ªFace Recognition from One Example View, Science, vol. 272, no. 5250, 1996.

[BRV02] Volker Blanz, Sami Romdhani, and Thomas Vetter, “Face identification across different poses and illuminations with a 3D morphable model,” in Proc. IEEE Inernational Conference on Automatic Face and Gesture Recognition, , pp. 202–207, 2002

[BS03] Bai, L. & Shen, L. , Combining wavelet and HMM for face recognition. In Proc. Of the 23rd Artificial Intelligence Conference (pp. 227-234), 2003.

[BS03] Bai, L. & Shen, L.. Face detection by orientation map matching. In Proc. of International Conference on Computational Intelligence for Modelling Control and Automation (pp. 363-370). 2003.

[BSXW07] Baochang Z., Shiguang S., Xilin C., and Wen G., “Histogram of Gabor Phase Patterns (HGPP):A Novel Object Representation Approach for Face Recognition”. IEEE Trans. on Image Processing, vol.16, No.1, pp57-68. 2007.

[BT02] L. M. Brown and Y.-L. Tan. Comparative study of coarse head pose estimation. Technical report, IBM T.J.Watson Research Center, Hawthorne, NY, 2002.

[BV03] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," IEEE PAMI, 25(9):1063-1074, 2003.

[BV99] V. Blanz and T. Vetter, “A morphable model for the synthesis of 3D faces,” in Proc. ACM SIGGRAPH, Mar. 1999, pp. 187–194.

[BWSXW06] Bingpeng M, Wenchao Z, Shiguang S, Xilin C, Wen G.,Robust Head Pose Estimation Using LGBP",ICPR, vol.2, pp512-515, 2006

[CCG07] X. Chai, S. Shan, X. Chen, and W. Gao. Locally linear regression for pose-invariant face recognition. IEEE Trans. Image Processing, 16(7):1716 – 1725, 2007.

[CET01] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[CLLH01] L-F. Chen, H-Y. Liao, J-C. Lin and C-C. Han, "Why recognition in a statistics-based face recognition system should be based on the pure face portion: a probabilistic decision-based proof", Pattern Recognition, Vol. 34, No. 7, pp. 1393-1403, 2001.

[CPT06] S. Cheng, S. Park, and M. Trivedi, "Multi-spectral and multiperspective video arrays for driver body tracking and activity analysis," Computer Vision and Image understanding, vol. 106, no. 2-3, 2006.

[CSB06] F. Cardinaux, C. Sanderson, and S. Bengio. User authentication via adapted statistical models of face images. *IEEE Trans. Signal Processing*, 54(1):361–373, 2006.

[CSB06] F. Cardinaux, C. Sanderson, and S. Bengio. User authentication via adapted statistical models of face images. *IEEE Trans. Signal Processing*, 54(1):361–373, 2006.

[CSCG07] Xiujuan Chai, Shiguang Shan, Xilin Chen and Wen Gao. Locally Linear Regression for Pose Invariant Face Recognition. IEEE Trans. on Image Processing. Vol.16, No.7, pp1716-1725, Jul. 2007.

[CT04] E. Murphy-Chutorian and M. Trivedi, "Robust real-time detection, tracking, and pose estimation of faces in video streams," in Proc. Int'l. Conf. Pattern Recognition, 2004, pp. 965–968.

[CT08] E. Murphy-Chutorian and M. Trivedi, "Hybrid head orientation and position estimation (HyHOPE): A system and evaluation for driver support," in Proc. IEEE Intelligent Vehicles Symposium, 2008.

[CT08b] E. Murphy-Chutorian and M. Trivedi,"Head Pose Estimation: A Survey. ", In IEEE PAMI, 2008.

[CT95] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham,"Active Shape Models - Their Training and Application",CVIU, 61(1): 38-59, 1995.

[CY04] D. Casasent and C. Yuan, “Face recognition with pose and illumination variations using new SVRDM supportvector machine,” Optical Engineering 43(8), pp. 1804-1813, August 2004.

[D02] Duin, R.P.W., The combining classifier: to train or not to train?, ICPR. (II: 765-770).2002.

[D93] Daugman, J.: High confidence visual recognition of persons by a test of statistical independence. IEEE Transactions on Pattern Analysis and Machine Intelligence 15 1148–1161, 1993.

[DBBB03] Draper, B. A., Baek, K., Bartlett, M. S., & Beveridge, J. R., Recognizing faces with PCA and ICA. Computer Vision and Image Understanding, 91, 115-137, 2003.

[DB04] M.B. Dias and B.F. Buxton, "Separating Shape and Pose Variations", IVC, 22(10): 851-861, 2004.

[DFB99] Duc, B., Fischer, S., & Bigun, J., Face authentication with Gabor information on deformable graphs. IEEE Transactions on Image Processing, 8, 504-516, 1999.

[DHS01] R.O. Duda, P.E.Hart and D.G.Stork, *Pattern Classification*, 2nd edition, Wiley Interscience, 2001.

[ECT98] G.J. Edwards, T.F. Cootes, and C.J. Taylor, “Face recognition using active appearance models,” in Proc. European Conference on Computer Vision, 1998, vol. 2, pp. 581–695.

[EMR00] S. Eickeler, S. M¨uller and G. Rigoll, “Recognition of JPEG Compressed Face Images Based on Statistical Methods”, Image and Vision Computing, Vol. 18, No. 4, pp. 279-287, 2000.

[EZ06] Everingham, M., & Zisserman, A. Regression and classification approaches to eye localization in face images. In international conference on automatic face and gesture recognition FG (pp. 441–446). 2006.

[FH07] Y. Fu and T. S. Huang, “hMouse: Head tracking driven virtual computer mouse,” in IEEE Workshop Applications of Computer Vision, 2007, pp. 30–35.

[FRVT02] FRVT 2002: Overview and summary, P Phillips, P Grother, R Micheals, D Blackburn, E Tabassi, J. Bone 2003 available at http://www.frvt.org/FRVT2002/documents.htm>.

[GA04] R. Gottumukkal and V. K. Asari, “An improved face recognition technique based on modular PCA approach,” Pattern Recognit. Lett., vol. 25, no. 4, pp. 429–436, 2004.

[GC94] A. Gee and R. Cipolla, "Estimating Gaze from a Single View of a Face", Proc. ICPR, pp. 758-760, 1994.

[GH04] Gründig M. and Hellwich,O., "3D Head Pose Estimation with Symmetry Based Illumination Model in Low Resolution Video", DAGM, LNCS, pp. 45-53, 2004.

[GIAA03] Gokberk, B., Irfanoglu, M. O., Akarun, L., & Alpaydin, E. (2003). Optimal Gabor Kernel Selection for Face Recognition. In Proceedings of the IEEE International Conference on Image Processing (pp. 677-680), 2003.

[GLK00] G. Guo, S. Z. Li, K. Chan, "Face Recognition by Support Vector Machines," p. 196, Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000.

[GLWH01] Y. Gao, M. K. H. Leung, W. Wang, and S. C. Hui, "Fast face identification under varying pose from a single 2-D model view," IEE Proc.-VISP, 148(4): 248-253, 2001.

[GMP00] Shaogang Gong, Stephen J McKenna, and Alexandra Psarrou. DYNAMIC VISION, From Images to Face Recognition. Imperial College Press, London, 2000.

[GMB04] Gross, R., Matthews, I., & Baker, S. (2004). Appearance-based face recognition and light-fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(4), 449–465.

[Gra78] Granlund, G. H., Search of A General Picture Processing Operator. Computer Graphics and Image Processing, 8, 155-173, 1978.

[GW93] R. C. Gonzales and R. E. Woods, Digital Image Processing, Addison-Wesley, Reading, Massachusetts, 1993.

[HCM05] Heusch, Cardinaux, Marcel,”Lighting Normalization Algorithms for Face verification”,IDIAP-COM 05-03, Mar. 2005

[HE02] Hesher, C. S. A. & Erlebacher, G. ,PCA of Range Images for Facial Recognition. In International Multiconference in Computer Science. 2002.

[HFT03] C. Hu, R. Feris, and M. Turk. Real-time view-based face alignment using active wavelet networks. International Workshop on Analysis and Modeling of Faces and Gestures, 2003.

[HLLYS04] Y. Huang, S. Lin, S. Z. Li, H. Lu, and H.-Y. Shum. Face alignment under variable illumination. In Proceedings of 6th IEEE International Conference on Automatic Face and Gesture Recognition, 2004.

[HRS08] Hayward, W. G., G. Rhodes and A. Schwaninger: An own-race advantage for components as well as configurations in face recognition. Cognition 106(2), 1017-1027 (02 2008)

[Hun08] Hung-Son Le, Face recognition-A single view based HMM approach. PhD Thesis. Umea University, Sweden, ISBN 978-91-7264-483-0, 2008.

[HW78]. Hubel, D., Wiesel, T.: Functional architecture of macaque monkey visual cortex. Proceedings of Royal Society on Biology 198 (1978) 1–59

[J06] Xiong Jiang, Ezra Rosen, Maximilian Riesenhuber, Thomas Zeffiro, and John VanMeter ,Volker blanz.: "Evaluation of a Shape-Based Model of Human Face Discrimination Using fMRI and Behavioral Techniques." Publishing in Neuron 50, 159-172, April 6, 2006.

[JDM00] A.K. Jain, R.P.W. Duin, and J. Mao, Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4-37, 2000.

[JV03] M. J. Jones and P. Viola. Fast multiview face detection. Technical report TR2003-96 MERL, 2003.

[K00] Kovesi, P.D., 2000 “Phase congruency: A low-level image invariant” Psychological Research,64, pp136-148.

[K02] L. I. Kuncheva, "A Theoretical Study on Six Classifier Fusion Strategies," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 2, pp. 281--286, 2002.

[K03] Kovesi,PD,2003,”phase congruency detects corners and edges”, in proc. Australian pattern recognition society conference, pp 309-318.

[K04] L.I.Kuncheva , Combining Pattern Classifiers: Methods and Algorithms. John Wiley and Sons. 2004, ISBN: 978-0-471-21078-8.

[Kan73] T. Kanade, Picture Processing by Computer Complex and Recognition of Human Faces, PhD thesis, Kyoto University, 1973.

[KISI06] M.S.Khalid, M.U.Ilyas, M.S.Sarfraz, M.A.Ijaz," Bhattacharyya Coefficient in Correlation of Gray Scale Objects", In Journal of Multimedia, Academy publishers, Vol. 1, No. 1, pp. 56-61, 2006.

[KK05] Kim, T., & Kittler, J. (2005). Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(3), 318–327.

[KL51] Kullback, S., and Leibler, R. A., , On information and sufficiency, Annals of Mathematical Statistics **22**: 79-86, 1951.

[KS02] V. Kruger and G. Sommer, "Gabor wavelet networks for efficient head pose Estimation," Image and Vision Computing, vol. 20, no. 9-10, pp. 665–672, 2002.

[KS05] M.S.Khalid, M.S.Sarfraz "Performance of a Similarity Measure in Grayscale Image Matching" IEEE International Conference of Emerging Technologies, ICET 2005, pp 121-125; Islamabad, Pakistan, 2005.

[KY03] Kanade, T., & Yamada, A. (2003). Multi-subregion based probabilistic approach towards pose-Invariant face recognition. In IEEE international symposium on computational intelligence in robotics automation (Vol. 2, pp. 954–959).

[L02] M. Lockie. Facial verification bureau launched by police it group. Biometric Technology Today, 10(3):3–4, 2002.

[L04] Lowe D.: Distinctive image features from scale-invariant keypoints. Int. Journal of computer vision, 2(60):91-110, 2004.

[L98]. Lindeberg T.: Feature detection with automatic scale selection. Int. Journal of computer vision, vol. 30 no. 2,pp 79-116, 1998.

[LC05] Liu, X., & Chen, T. (2005). Pose-robust face recognition using geometry assisted probabilistic modeling. In International conference on computer vision and pattern recognition (CVPR) (Vol. 1, pp. 502–509).

[LC06] S. Lucey and T. Chen, "Learning Patch Dependencies for Improved Pose Mismatched Face Verification," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 17-22, 2006.

[LC08] S. Lucey and T. Chen. A viewpoint invariant, sparsely registered, patch based, face verifier. IJCV, January 2008.

[LG97] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back. “Face recognition: A convolutional neural network approach”, IEEE Trans. Neural Networks, 8:98–113, 1997.

[Liu04]. Liu, C.: Gabor-based kernel PCA with fractional power polynomial models for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 26 ,pp 572–581, 2004.

[LK06] H.S. Lee , D. Kim, Generating frontal view face image for pose invariant face recognition, PR letters vol.27, No. 7, pp 747-754, 2006.

[LR03] Lee, M. W. & Ranganath, S. (2003). Pose-invariant face recognition using a 3D deformable model. Pattern Recognition, 36, 1835-1846.

[LR03] Lee, M. W. & Ranganath, S. (2003). Pose-invariant face recognition using a 3D deformable model. Pattern Recognition, 36, 1835-1846.

[LTC97] Lanitis, A., Taylor, C. J., & Cootes, T. F. (1997). Automatic interpretation and coding of face images using flexible models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19, 743-756.

[LVBM93]. Lades, M., Vorbruggen, J., Budmann, J., Lange, J., Malsburg, C., Wurtz, R.: Distortion invariant object recognition on the dynamic link architecture. IEEE Transactions on Computers 42 (1993) 300–311

[LW02] Liu, C., Wechsler, H.: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image Processing 11 ,pp-467–476, 2002.

[M02] Martınez A. M.: Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE Trans. PAMI, vol. 24, no. 6, pp. 748–763, 2002.

[M05] E. Learned-Miller. Data driven image models through continuous joint alignment. PAMI, 2005.

[MB04] I. Matthews and S. Baker. Active appearance models revisited. International Journal of Computer Vision, 60(2):135 164, Nov. 2004.

[MCVC06] Picozzi, Marta., Macchi Cassia, Viola. and turati, chiara. "The development of configural face processing: the face inversion effect in preschool-aged children" Annual meeting of the XVth Biennial

[MGC96] S.McKenna, S. Gong and J.J. Collins. Face Tracking and Pose Representation. British Machine Vision Conference, Edinburgh, Scotland, 1996.

[MH03] Mu, X. Y. & Hassoun, M. H., Combining Gabor features: Summing vs.voting in human face recognition. 2003 IEEE International Conference on Systems, Man and Cybernetics, Vols 1-5, Conference Proceedings, 737-743, 2003.

[MK01] A.M. Martinez and A.C. Kak, "PCA versus LDA", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 228-233,. 2001.

[MM99] Menser, B. & Muller, F., Face detection in color images using principal components analysis. In Proc.of Seventh International Conference on Image Processing and its Applications (pp. 620-624),1999.

[MN95] H. Murase and S.K. Nayar. Visual Learning and Recognition of 3-D Objects form Appearance. International Journal of Computer Vision, 14:5-24,1995.

[MO87] Morrone, M.C., Owens, R.A., 1987 "Feature detection from local energy". PR Letters(6), pp 303-313.

[MP97] Moghaddam B. and Pentland A.: Probabilistic visual learning for object recognition. IEEE Trans. PAMI, vol. 19, no. 7, pp. 696–710, 1997.

[MSLD07] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell, "Head gestures for perceptual interfaces: The role of context in improving recognition," Artificial Intelligence, vol. 171, no. 8–9, pp. 568–585, 2007.

[MS05]. Mikolajczyk and Schmid C.: Performance evaluation of local descriptors. PAMI,27(10):31-47, 2005.

[MT08] Michaeli, Tomer, Face normalization for recognition and enrollment. Patent Number EP1872303, January 2008

[MWB97] Morris Moscovitch , Gordon Winocur , Marlene Behrmann ,Society For Neuroscience. "Facing The Issue: New Research Shows That The Brain Processes Faces And Objects In Separate Brain Systems." Journal of Cognitive Neuroscience ,Sept 1997.

[NH99] Nefian, A. & Hayes, M., An embedded hmm-based approach for face detection and recognition. In Proc.of IEEE Int. Conf. On Acoustics, Speech and Signal Processing (pp. 3553-3556), 1999.

[OBRG04] J. Ortega-Garcia, J. Big¨un, D. Reynolds, and J. Gonzales-Rodriguez. Authentication gets personal with biometrics. IEEE Signal Processing Magazine, 21(2):50–62, 2004.

[OF96]. Olshausen, B., Field, D.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381 (607–609, 1996.

[OPM02] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, Jul. 2002.

[PC05] R.Patnaik and D. P. Casasent, 2005 "MINACE-filter-based facial pose estimation" in Biometric Technology for Human Identification., Proc SPIE, pp 460-467.

[PC97] C. Padgett and G. Cottrell. Representing face images for emotion classification. Advances in Neural Information Processing Systems, 9, 1997.

[PEWF08] S. J. D. Prince, J. H. Elder, J. Warrell, and F. M. Felisberti. Tied factor analysis for face recognition across large pose differences. IEEE Trans. PAMI, 30(6):970–984, 2008.

[PFSBW06] P. J. Phillips, P. J. Flynn, W. T. Scruggs, K. W. Bowyer, and W. Worek, "Preliminary face recognition grand challenge results," in Seventh International Conference on Automatic Face and Gesture Recognition, 2006, pp. 15–24.

[PMRR00] P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss. The FERET evaluation methodology for face-recognition algorithms.IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(10):1090–1104, 2000.

[PMS94] A. Pentland, B. Moghaddam, and T. Starner, "View-Based and modular eigenspaces for face recognition," in Proc. IEEE Conf. Compute Vision and Pattern Recognition, pp. 84–91, 1994.

[PMS94] A. Pentland, B. Moghaddam and T. Starner. Viewbased and Modular Eigenspaces for Face Recognition.IEEE CVPR, pp. 84-91, 1994.

[RB95]. Rao, R., Ballard, D.: An active vision architecture based on iconic representations. Artificial Intelligence 78, 461–505,1995.

[RBK98] H. Rowley, S. Baluja, and T. Kanade. Neural networkbased face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(1):2338, Jan. 1998.

[RBK98] H. A. Rowley, S. Baluja, and T.Kanade. 1998 “Neural network-based face detection” IEEE PAMI, 20(1):23–38.

[RBV02] S. Romdhani, V. Blanz and T.Vetter, “Face identification by fitting a 3D morphable model using linear shape and texture error functions,” ECCV, 2002.

[RCBM06] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariethoz. Measuring the performance of face localization systems. *Image and Vision Computing*, 24:882–893, 2006.

[RO97] Robbins, B., Owens, R. 1997 “2D feature detection via local energy” Image and Vision Computing 15,pp353-368.

[RTG00] Rubner, Y., Tomasi, C., and Guibas, L. J., 2000. The Earth Mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, **40**(2): 99-121.

[S92] Silverman BW.: Density estimation for statistics and data analysis. Chapman and Hall. 1992.

[SC00] Schiele, B., Crowley, J.: Recognition without correspondence using multidimensional receptive field histograms. International Journal on Computer Vision 36 31–52, 2000.

[SGCCY04] Shiguang Shan, Wen Gao1, Yizheng Chang, Bo Cao, Pang Yang“Review the Strength of Gabor features for Face Recognition from the Angle of its Robustness to Mis-alignment” ICPR, 2004.

[SGO99] J. Sherrah, S. Gong, and E.-J. Ong, “Understanding pose discrimination in similarity space,” in British Machine Vision Conference, 1999, pp. 523–532.

[SH97] F. S. Samaria and A. C. Harter. “Parameterization of a stochastic model for human face identification” 2nd IEEE workshop on Applications of Computer Vision,1994.

[SH07] M. S. Sarfraz, O. Hellwich.:"Robust Facial Pose Estimation for Face Recognition", In Proceedings of 7th IAPR Open German Russian Workshop on Pattern Recognition and Image Analysis, Ettlingen, Germany, August 2007.

[SH08a] M.S.Sarfraz, O.Hellwich,” An efficient front-end facial Pose estimation system for Face Recognition”, Int. Journal of Pattern Recognition and Image Analysis, dist. by Springer, Vol.18(3) pp. pp. 434–441. 2008.

[SH08b] M.S.Sarfraz and O. Hellwich “Statistical Appearance Models for Automatic Pose Invariant Face Recognition.” 8th IEEE Int. conference on Face and Gesture recognition, FG September 2008.

[SH08c] M.S.Sarfraz, O.Hellwich ,"Head Pose Estimation in Face Recognition across Pose Scenarios" in Int. conference on computer vision theory and applications VISAPP, January 2008, Vol. 1, pp. 235 -242, Portugal.

[SH08d] M.S.Sarfraz , O.Hellwich, "Learning Probablistic Models for Recognizing faces under Pose Variations" in VISAPP-IMTA Workshop Image mining, theory and applications, January 2008, pp. 122-132, Portugal.

[SH09a] M.S.Sarfraz and O.Hellwich, “On Head Pose Estimation in Face Recognition”, J. Braz, A. Ranchordas, H. Araújo and J. Jorge Editors, Advances in Computer Graphics and Computer Vision. Springer, 2009. (To appear).

[SJH06] M.S.Sarfraz, M.Jäger, O.Hellwich “Performance Analysis of Classifiers on Face Recognition”, 5th IEEE Advances in Cybernetics Systems AICS conference, United Kingdom, pp. 255 -264, 2006.

[SK02] C. A. Shipp and L.I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers," International Journal of Information Fusion, vol.3, no. 2, pp. 135--148, 2002.

[SN05] J. Ruiz-del-Solar, P. Navarrete, Eigenspace-based face recognition: a comparative study of different approaches, IEEE Transactions on Systems, Man and Cybernetics, Part C, Vol. 35, Issue 3, August 2005, pp. 315-325.

[SS02] Bernhard Schölkopf and A. J. Smola: *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[SS99] R. Schapire and Y. Singer, “Improved boosting algorithms using confidence-related predictions”. Machine Learning, 37(3), pp.297-336, 1999

[SWCG06] Schwaninger, A., C. Wallraven, D. W. Cunningham and S. Chiller-Glaus: Processing of identity and emotion in faces: a psychophysical, physiological and computational perspective. Progress in Brain Research 156, 321-343 (09 2006)

[SY94] Samaria, F. & Young, S., Hmm-Based Architecture for Face Identification. Image and Vision Computing, 12, 537-543, 1994.

[TC05] K. Tan and S. Chen, Adaptively weighted sub-pattern PCA for face recognition, Neurocomputing 64, pp. 505–511, 2005.

[TCZZ06] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang. Face recognition from a single image per person: A survey. Pattern Recogn., 39(9):1725–1745, 2006.

[TP91] M. Turk and A. Pentland, “Eigenfaces for Recognition”, Journal of Cognitive Neuroscience, Vol. 3, No. 1, pp. 71-86. 1991

[UB97] S. Ullman and R. Basri, "Recognition by Linear Combinations of Models", IEEE PAMI, 13: 992-1006, 1991.

[UVS02] Ullman, S., Vidal-Naquet, M., and Sali, E., Visual features of intermediate complexity and their use in classification. Nature Neuroscience, (7):682–687, 2002.

[VA96] D. Valentin and H. Abdi. Can a Linear Autoassociator Recognize Faces From New Orientations. Journal of the Optical Society of America A-Optics, Image Science and vision, 13(4), pp. 717-724, 1996.

[VT02] Vasilescu MA.O. and Terzopoulos D.: Multilinear analysis of image ensembles: TensorFaces. ECCV, vol. 2350, pp. 447–460, 2002.

[W02] A.R.Webb, Statistical pattern recognition, second edition, john willey & sons, 2002.

[W07] L. Wasserman, *All of Nonparametric Statistics*. 3rd edition, Springer, 2007.

[WCH02] Wang, Y. J., Chua, C. S., & Ho, Y. K. Facial feature detection and face recognition from 2D and 3D images. Pattern Recognition Letters, 23, 1191-1202, 2002.

[WFKM97]. Wiskott, L., Fellous, J., Kr¨uger, N., Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 775–779.

[WFT01] Y.Wei, L.Fradet, and T.Tan. Head pose estimation using gabor eigenspace modeling. Technical report, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, 2001. 7.

[WQ02] Wang, X. L. & Qi, H. R., Face Recognition Using Optimal Nonorthogonal Wavelet Basis Evaluated by Information Complexity. In

Proc. of the 16th Internatinal Conference on Pattern Recognition (pp. 164-167), 2002.

[WTJ06] P. Wang, L. C. Tran, and Q. Ji. Improving face recognition by online image alignment. Pattern Recognition, 2006.

[WYS02] Wu, H. Y., Yoshida Y., & Shioyama, T., Optimal Gabor filters for high speed face identification. In Proc. of Int. Conf. on Pattern Recognition (pp. 107-110), 2002.

[YC05] C. Yuan and D. P. Casasent,2005 "Face recognition and verification with pose and illumination variations and imposer rejection," in Biometric Technology for Human Identification., Proc SPIE.

[YKA02] M.-H. Yang, D.J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(1):34–58, Jan 2002.

[ZC00] Zhao, W. & Chellapa, R. , 3D model enhanced face recognition. In Proc. International Conference on Image Processing (pp. 50-53), 2000.

[ZC00] Zhao,W., Chellappa, R. (2000). SFS based view synthesis for robust face recognition. In International conference on automatic face and gesture recognition (pp. 285–292).

[ZCRP03] W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips, Face Recognition: A Literature Survey, ACM Computing Surveys, 2003, pp. 399-458

[ZCSCß7] G. Zhao, L. Chen, J. Song, and G. Chen, "Large head movement tracking using SIFT-based registration," in Proc. Int'l. Conf. Multimedia, 2007, pp. 807–810.

[ZG06] S. Zhao and Y. Gao, 2006 "Automated Face Pose Estimation Using Elastic Energy Models", The 18th ICPR, pp.618-621, Vol. 4.

[ZJN07] Jie Zou; Qiang Ji; Nagy, G., A Comparative Study of Local Matching Approach for Face Recognition, IEEE Transactions on Image Processing, Volume 16, Issue 10, pp :2617 – 2628, 2007.

[ZJN07] Jie Zou, , Qiang Ji, and George Nagy, "A Comparative Study of Local Matching Approach for Face Recognition" IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 16, NO. 10, OCTOBER 2007.

[ZKCSW98] Zhao, W., Krishnaswamy, A., Chellapa, R., Swets, D. L., & Weng, J. J.,

Discriminant Analysis of Principal Components for Face Recognition. In H.Wechsler, P.J. Phillips, V. Bruce, F. F. Soulie, & Y. P. Huang (Eds.), Face Recognition: From Theory to Applications (CAR-TR-914 ed., pp. 73-85). Springer-Verlag, 1998.

[ZLG05] Zhang, Lingyun and Garrison W. Cottrell, Holistic processing develops because it is good. In Proceedings of the 27th Annual Cognitive Science Conference, Italy, 2005.

[ZSGCZ05] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local Gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition," in Proc. ICCV, pp. 786–791. 2005.

[ZZTS05] Y. Zhou, W. Zhang, X. Tang, and H. Shum. A Bayesian mixture model for multi-view face alignment. CVPR, 2005.